Μέθοδοι Ανάλυσης Κυτταρολογικών Εικόνων

Η ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

υποβάλλεται στην
ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης
του Τμήματος Πληροφορικής Εξεταστική Επιτροπή

από τη

Μαρίνα Ε. Πλησίτη

ως μέρος των Υποχρεώσεων για τη λήψη του

ΔΙΔΑΚΤΟΡΙΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

Ιανουάριος 2012

Τριμελής Συμβουλευτική Επιτροπή

Χριστόφορος Νίκου, Επίκουρος Καθηγητής Τμήματος Πληροφορικής του Παν/μίου Ιωαννίνων (Επιβλέπων)

Λυσίμαχος-Παύλος Κόντης, Επίκουρος Καθηγητής Τμήματος Πληροφορικής του Παν/μίου Ιωαννίνων

Αριστείδης Λύκας, Αναπληρωτής Καθηγητής Τμήματος Πληροφορικής του Παν/μίου Ιωαννίνων

Επταμελής Εξεταστική Επιτροπή

Χριστόφορος Νίκου, Επίκουρος Καθηγητής Τμήματος Πληροφορικής του Παν/μίου Ιωαννίνων (Επιβλέπων)

Λυσίμαχος-Παύλος Κόντης, Επίκουρος Καθηγητής Τμήματος Πληροφορικής του Παν/μίου Ιωαννίνων

Αριστείδης Λύκας, Αναπληρωτής Καθηγητής Τμήματος Πληροφορικής του Παν/μίου Ιωαννίνων

Γεώργιος Τσιχριντζής, Καθηγητής Τμήματος Πληροφορικής του Πανεπιστημίου Πειραιώς

Θεόδωρος Βλάχος, Αναπληρωτής Καθηγητής Τμήματος Τεχνών Ήχου και Εικόνας του Ιονίου Παν/μιου

Γεώργιος Μανής, Επίκουρος Καθηγητής Τμήματος Πληροφορικής του Παν/μίου Ιωαννίνων

Κων/νος Μπλέκας, Επίκουρος Καθηγητής Τμήματος Πληροφορικής του Παν/μίου Ιωαννίνων

Αφιερωση

Στους ανθρώπους που με πίστεψαν, με στήριξαν και με ενέπνευσαν.

Σε αυτούς που η πολύτιμη συμβολή τους ήταν καθοριστική
προϋπόθεση για την ολοκλήρωση αυτής της διατριβής.

Στον επιβλέποντα καθηγητή μου κ. Χριστόφορο Νίκου

στους γονείς μου Ευάγγελο και Βασιλική

και φυσικά

στο σύζυγό μου Γιώργο

και στα παιδιά μου

Αγγελική και Παναγιώτη ...

που αποτελούν για μένα αστείρευτη πηγή χαράς, αγάπης και δύναμης.

Ευχαριστιες

Επειδή πιστεύω πως κάθε έργο εκπορεύεται, πέραν από τον δημιουργό του,
και από όλους εκείνους που σε αμφίδρομη σχέση μαζί του, του δίνουν
τη δυνατότητα να καταστεί δημιουργός, οι ευχαριστίες μου προς όλους αυτούς
ελάχιστο αντίδωρο για τη δυνατότητα που μου πρόσφεραν.

Πρώτον απ' όλους ευχαριστώ τον επιβλέποντα καθηγητή μου κ. Χριστόφορο Νίκου, Επίκουρο Καθηγητή Τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων, στον οποίο οφείλω την πραγμάτωση της συγκεκριμένης διδακτορικής διατριβής. Τον ευχαριστώ τόσο για την εμπιστοσύνη που μου έδειξε αναλαμβάνοντας την επίβλεψη της διατριβής μου, όσο και για την άψογη συνεργασία μας, την αμέριστη βοήθειά του σε κάθε πρόβλημα που αντιμετωπίσαμε, την στήριξή του σε κάθε δυσκολία που προέκυπτε, την κατανόησή του, την υπομονή του, τις γνώσεις και τις κατευθύνσεις που μου πρόσφερε απλόχερα, έτσι ώστε αυτή η διδακτορική διατριβή να ολοκληρωθεί με επιτυχία. Η συνεργασία μου με τον κ. Νίκου υπήρξε εμπειρία ζωής, αφού εκτός από τις γνώσεις που απέκτησα, βίωσα το πώς ορίζεται ο ακαδημαϊκός δάσκαλος, ο άξιος επιστήμονας και το ακαδημαϊκό ήθος στην πραγματική τους διάσταση, έννοιες που ευρέως απαξιώνονται από πολλούς σήμερα.

Ένα μεγάλο ευχαριστώ οφείλω στην οικογένειά μου, το σιωπηλό σημείο αναφοράς μου, που πάντα είναι δίπλα μου και με στηρίζει σε όλες τις επιλογές μου. Ευχαριστώ το σύζυγό μου Γιώργο για την αγάπη του, την κατανόησή του, τη θετική του διάθεση, την ηθική συμπαράσταση και την ικανότητά του να με εκπλήσσει ευχάριστα σε κάθε αντιξοότητα. Ευχαριστώ τα παιδιά μου, Αγγελική και Παναγιώτη, που αποτελούν πηγή έμπνευσης και εφαλτήρια δύναμη στη ζωή μου, εκφρασμένη και στην υπόψη διατριβή, για το χρόνο που μου δάνεισαν, για τη χαρά και την αγάπη που μου προσφέρουν και για τις αξέχαστες εμπειρίες που απολαμβάνω καθημερινά.

Ευχαριστώ τους γονείς μου, Ευάγγελο και Βασιλική, για την υποστήριξη και την υπομονή τους, γιατί χωρίς αυτή την πολύτιμη βοήθειά τους, η διδακτορική διατριβή δεν θα μπορούσε να ολοκληρωθεί. Είναι αυτοί οι αφανείς ήρωες, που πάντα βρίσκονται κοντά μου και με αγάπη και διακριτικότητα δίνουν απεριόριστα την έμπρακτη στήριξή τους.

Θα ήθελα επίσης να ευχαριστήσω όλους τους ανθρώπους που με το δικό τους τρόπο ο καθένας βοήθησαν, κατά την εκπόνηση αυτής της διατριβής. Ευχαριστώ τους γονείς του συζύγου μου, Παναγιώτη και Αγγελική Γκανιάτσα, για τη βοήθεια και τη στήριξή τους. Ευχαριστώ τους συγγενείς μου Κωνσταντίνο και Ελένη Μπουκουβάλα, για τις συμβουλές, τη συμπαράσταση και τη βοήθειά τους. Ευχαριστώ τα αδέρφια μου, Κώστα και Βίκη, Γιώργο και Μάχη, και τον Γιάννη, για τη διαρκή ηθική συμπαράσταση, καθώς και την υπ. διδ. Ευανθία Τριπολίτη, για την στήριξη και ουσιαστική βοήθεια σε θέματα ερευνητικά και μη. Ευχαριστώ την αδερφική μου φίλη Σπυριδούλα Σιχλιμίρη, που με συνόδευσε στο ταξίδι μου στη Βαλένσια και που ήταν και είναι πάντα δίπλα μου. Ευχαριστώ τον καλό φίλο Χρίστο Παλαιοπάνο για τις υποδείξεις και διορθώσεις. Ευχαριστώ τη φίλη μου Τατιάνα Παπαδοπούλου, για τις εύστοχες συμβουλές που μου έδωσε σε καίριες στιγμές, και που τελικά καθόρισαν την πορεία της διατριβής μου. Ευχαριστώ τους υπ. διδ. Δημήτρη Γερογιάννη και Μιχάλη Βρίγκα, για τις ευχάριστες στιγμές που περάσαμε στο ίδιο εργαστήριο. Επίσης, ευχαριστώ την κ. Αντωνία Χαρχαντή, Επίκουρη Καθηγήτρια Ιατρικής Σχολής του Πανεπιστημίου Ιωαννίνων για τις κατευθύνσεις που μας έδωσε, οι οποίες ήταν απαραίτητες για την ολοκλήρωση της διατριβής.

Μαρίνα Ε. Πλησίτη
Ιανουάριος 2012

Table of Contents

1 Introduction 1

1.1 Thesis Contribution . . . 3

2 A review of cervical cell image analysis methods 6

2.1 Pap smear images . . . 6

2.2 Segmentation of Pap smear images . . . 8

2.2.1 Thresholding . . . 9

2.2.2 Edge Detection . . . 10

2.2.3 Mathematical morphology . . . 11

2.2.4 Pixel classification . . . 11

2.2.5 Template matching . . . 12

2.2.6 Deformable models . . . 13

2.2.7 Segmentation of overlapping nuclei in cervical images . . . 13

2.3 Classification of Pap smear Images . . . 14

2.3.1 Features of cervical cells used for classification . . . 15

2.3.2 Classification techniques . . . 16

2.4 Conclusion . . . 17

3 Automated Detection of Cell Nuclei using Morphological Reconstruction and Clustering 18

3.1 Introduction . . . 18

3.2 Methodology . . . 21

3.2.1 Preprocessing . . . 21

3.2.2 Detection of candidate cell nuclei centroids . . . 23

3.2.3 Refinement of candidate cell nuclei centroids . . . 27

3.2.4 Decision . . . 29

3.3 Results . . . 31

3.3.1 Study Group . . . 31

3.3.2 Numerical evaluation . . . 32

3.4 Discussion . . . 34

3.4.1 Evaluation of the proposed method . . . 34

3.4.2 Comparison of the proposed method with other methods . . . 36

3.5 Conclusion . . . 38

4 Cell Nuclei Extraction by Combining Shape, Texture and Intensity Features 46

4.1 Introduction . . . 46

4.2 Methodology . . . 48

4.2.1 Detection of the nuclei and cytoplasm markers . . . 48

4.2.2 Morphological color gradient image . . . 49

4.2.3 Clustering of the candidate nuclei . . . 49

4.3 Results . . . 57

4.3.1 Study group . . . 57

4.3.2 Numerical evaluation . . . 58

4.4 Discussion . . . 66

4.5 Conclusion . . . 69

5 Overlapping Cell Nuclei Segmentation Using A Spatially Adaptive Active Physical Model 70

5.1 Introduction . . . 70

5.2 Methodology . . . 73

5.2.1 Training phase . . . 73

5.2.2 Segmentation of overlapping nuclei boundaries . . . 77

5.3 Experimental results and discussion . . . 84

5.3.1 Study group . . . 84

5.3.2 Numerical evaluation . . . 85

5.3.3 Relation between the weight values and the overlapping percentage 93

5.4 Conclusion . . . 95

6 Cervical Cell Classification Based Exclusively On Nucleus Features 96

6.1 Introduction . . . 96

6.2 Methodology . . . 98

6.2.1 Study group . . . 98

6.2.2 Feature generation and dimensionality reduction . . . 98

6.3 Results and discussion . . . 101

6.4 Conclusion . . . 102

7 Conclusions and Perspectives 104

List of Figures

1.1 The steps of the Pap test. (a) The sample of cells is collected from the cervix using a small cone-shaped brush, (b) the material is transferred from sampling instrument to the slide, (c) a typical Pap smear slide, (d) examination of the slide under the microscope for the detection of abnormalities. . . . 2

2.1 A conventional Pap smear image. 7

2.2 The several types of cells that are included in Pap smear images (see text for details). In this image, the high degree of cell overlapping, the inhomogeneity in cells intensity and the existence of many artifacts are clearly present. 8

2.3 (a) Original cell image and (b) the structural parts of the cell; the nucleus and the cytoplasm. 9

2.4 Schema of Pap smear microscopic image classification process. 15

3.1 (a) Initial cell image. (b) Mapping of the intensity values in the color space, where high intensity values are represented by red color and small intensity values are represented by blue color. Point A corresponds to the location of a true nucleus and points B and C correspond to areas of cell overlapping. (c) Mapping of the initial image in 3D space. The points A, B and C are lying in the same intensity level but only point A corresponds to the location of a true nucleus. As it is observed, the local depth h_A of this point is more pronounced with respect to h_B and h_C. 19

3.2 Schema of the proposed method. 20

3.3 (a) The initial Pap smear image, and (b) the binary mask which is obtained after the preprocessing step. 24

3.4 (a) Initial image of a cell cluster with overlapped cells, (b) the resulted image with the suppressed regional minima, (c) the areas of regional minima (notice that both cell nuclei and non-cell nuclei are extracted at this step) and (d) the centroids of the areas of regional minima. 26

3.5 Illustration of the different steps of the refinement procedure. (a) Initial image, (b) eroded image, (c) filtered image, (d) difference of images in (b) and (c), (e) contrast enhanced image, (f) construction of the search space, (g) determination of pixels in the nucleus circumference by selecting the local maxima of the gradient amplitude, (h) the resulted nucleus contour superimposed onto initial image and (i) the initial (black cross) and the refined (white circle) centroid of the nucleus. 28

3.6 The detection of regional minima of the image results in many candidate points. This is also true for a single nucleus. 30

3.7 (a) The initial image with the centroids of regional minima depicted with an "×". (b) The result of the application of the distance depended rule, (c) the result of the application of the FCM, where the positive class (nuclei centroids) is depicted with a cross and the negative class (other findings) is depicted with a circle. (d) Resulted centroids of the positive class. 40

3.8 Results of the application of the FCM clustering with respect to (a) *Sensitivity* and (b) *Specificity* (see text for the description of the pattern sets). 41

3.9 Results of the application of the SVM classification with respect to (a) *Sensitivity* and (b) *Specificity* (see text for the description of the pattern sets). 42

3.10 ROC curve used for the selection of the distance threshold in the distance dependent rule. Notice that for threshold values higher than 8, true nuclei centroids are missed. 43

3.11 Rate of the true nuclei detected (true positives) and the false nuclei centroids detected (false positive) for different thresholds in regional minima depth. Notice that for the threshold value of 15 we obtain the maximum number of true nuclei centroids. 43

3.12 Comparative results of our method and the pixel classification schemes proposed in [1] in terms of correct nuclei localization. The index of performance is the number of connected components and the number of true nuclei (true positives) they contain. An algorithm performs well if each connected component it produces contains only one nucleus. The vertical axis has a logarithmic scale. 44

3.13 a) The initial image, (b) the result of the preprocessing step (denoted with the black line) in the color image, (c) the result of the preprocessing step in the grayscale image. The missed nuclei are marked with the arrows, (d) a part of the initial image, (e) the result of the detection of regional minima step (denoted with white lines) in the color image, (f) the result of the same step in the grayscale image. The missed nuclei of this step are marked with the arrows. 45

4.1 (a) Initial image of overlapped cells, (b) the detected nuclei markers, (c) the corresponding color morphological gradient image, (d) the watershed segmentation. 48

4.2 (a) Initial image of overlapped cells and (b) the corresponding grayscale image, in which we apply the Canny edge detector. Using a small threshold results in (c) an image with many undesired edges, while using a high threshold results in (d) an image with several significant edges missing. . . 50

4.3 The detected centroids of the regional minima in the image. The true nuclei locations are represented by a yellow cross and the false positive findings are represented by a black circle. 51

4.4 (a)-(b) The result of the watershed transform in parts of two different cell images. The regions R_1 and R_2 that are detected in both images with the watershed transform are joined with a line for better visualization purposes. In (a), the detected areas R_1 and R_2 correspond to the areas of true nuclei, while in (b), the detected area R_1 corresponds to a nucleus and the area R_2 corresponds to a cytoplasm overlapping area. The variation of the average color image intensity value along the line which joins the areas R_1 and R_2 is depicted in (c). Notice that for the area R_1 we observe sharp reduction of the intensity value in both images. For the area R_2, although the average intensity value is similar in both images, sharper intensity reduction (in relation with its neighborhood pixels) occurs only for the true nucleus in image (a). This indicates that the use of the neighborhood of each detected area contributes in the recognition of the true nuclei. 52

4.5 The selected areas for the construction of the feature set. (a) A cell from the initial image, (b) the detected nucleus boundary with the watershed transform and the enclosed area A, (c) the area B of the bounding box of the detected boundary, (d) the area of the neighborhood Ngh ($A^c \cap B$) of the detected nucleus. 53

4.6 The topology of the neighborhood used for the calculation of the LBP [2, 3]: (a) circle, (b) hyperbola. 53

4.7 Representative histograms of some features of the watershed and the GVF segmentation. Notice that their distribution consists of a single blob and this allows their discretization into three states at the positions $\mu \pm \sigma$. . . 56

4.8 The leave-one-out and global mRMR feature rank for the watershed transform. For the leave-one-out mRMR feature rank the standard deviation is also depicted with error bars. 59

4.9 The leave-one-out and global mRMR feature rank for the ACM segmentation algorithm. For the leave-one-out mRMR feature rank the standard deviation is also depicted with error bars. 60

4.10 The leave-one-out and global mRMR feature rank for the GVF segmentation algorithm. For the leave-one-out mRMR feature rank the standard deviation is also depicted with error bars. 60

4.11 Results in terms of the HM measure for the K-means clustering for ACM, GVF and watershed segmentation for both (a) global and (b) leave-one-out mRMR rank. The vertical line indicates the number of features where the HM measure takes its maximum value for the three segmentation methods. These values of HM are contained in Table 4.5. 63

4.12 Results in terms of the HM measure for spectral clustering for ACM, GVF and watershed segmentation for both (a) global and (b) leave-one-out mRMR rank. The vertical line indicates the number of features where the HM measure takes its maximum value for the three segmentation methods. These values of HM are contained in Table 4.5. 64

4.13 Results in terms of the HM measure for the SVM classification for ACM, GVF and watershed segmentation for both (a) global and (b) leave-one-out mRMR rank. For comparison purposes, the indicative values for HM measure were evaluated using the first 26 features. The values of HM are contained in Table 4.5. 65

4.14 (a)-(c) Segmentation results for several detected nuclei. 66

4.15 Representative cases of failure for ACM and GVF segmentation in images with (a) weak gradient at the nucleus boundary, (b) the inhomogeneities of the nucleus intensity and (c) the existence of high value of gradient in the neighborhood of the nucleus boundary. 68

5.1 Convergence of the physical model to the boundary of a cell nucleus in the training phase. (a) The initial image, (b) the manually traced nucleus boundary, (c) registration of the boundary to a reference shape oriented horizontally, (d) the distance transform of (c), (e) the initial (red) and the final (white) position of the physical model, (f) the detected nucleus boundary in the initial image rotated appropriately. The figure is better seen in colour. 76

5.2 Modes of variation of the nuclei shape. Each row corresponds to the variations of the shape of the model by approximating $\tilde{\mathbf{U}}$ using only one eigenvector $\mathbf{a}_j$, $j = 1, ...4$ and setting the corresponded value of b_j to the values in the last row, with all the values of $b_k = 0$, for $k \neq j$ in (5.19). In the middle column, the mean shape is depicted. 78

5.3 The basic steps of the segmentation of the overlapping nuclei (see text for details). 78

5.4 (a) The initial image, (b) image obtained after noise reduction, histogram equalization and H-minima transform of the greyscale counterpart of the initial image, (c) the result of the application of the Canny edge detector in (b), (d) initial placement of the learnt model on each nucleus, after the chamfer matching. The figure is better seen in colour. 80

5.5 Initial positions of the models in several images. Notice that best initial positions are obtained with the rotation and the scaling of the learnt model. The figure is better seen in colour. 81

5.6 (a) Initial image and, (b) initial position of the two models. Result obtained using (c) small and (d) large weights values for the image force of one of the two nuclei. Notice that in (c) there exist small differences from the initial position of the models. However in (d), the model of one nucleus converges in a position of high gradient of the image, resulting in the erroneous identification of the boundary of the area containing both of the nuclei. The figure is better seen in colour. 83

5.7 (a) Initial image, (b) initial position of the models, (c) calculation of the overlapping area, (d) points lying in the area of overlap, denoted with yellow, have very small weights with respect to the rest of the points. The figure is better seen in colour. 84

5.8 The result of the proposed method using the same value for the image force term, for all of the points of the model. Notice that best results are obtained using the value of 0.1 and the corresponding Hausdorff distance is 24.29. 86

5.9 The result of the proposed method using the same value $w_i = 0.1$ for the image force term in the points lying in the non-overlapping area of the nuclei, and multiple values $w_{1,2,\ldots,k}$ for the k points of the model lying in the overlapping area. Notice that best results are obtained using the value of 0.001 and the corresponding Hausdorff distance is 19.58. 87

5.10 Segmentation results of the method. (a) Ground truth, (b) nuclei boundaries obtained with the convergence of the models. The figure is better seen in colour. 88

5.11 (a) Representative images containing overlapped nuclei. (b) Edge images resulted after the application of the Canny edge detector (initialization step). The point A in the first image indicates some falsely detected edges in the inner area of the nucleus. The points A and B in the second image indicate the existence of discontinuities. (c) The corresponding distance transform. 89

5.12 Segmentation results. (a) The ground truth in the initial image, the segmentation results for (b) the proposed method, (c) the ellipse fitting algorithm, (d) the standard ASM algorithm, (e) H-minima marker extraction [4], (f) Bayesian [5]. 90

5.13 The original images of the training set used for the construction of the synthetic image. 93
5.14 Construction of a synthetic image. The degree of area overlapping for the upper left nucleus is 10%, and for the bottom right is 90%. 94
5.15 The Hausdorff distance in the synthetic image with respect to the nuclei overlapping percentage. The curves correspond to the segmentation obtained using the specific rates of the weight values of the non-overlapping areas to the weight values of the overlapping areas. 95

6.1 (a) An isolated cell and (b) overlapping cells. Notice that the cytoplasm area is clearly recognized in (a) while in (b) its determination is very ambiguous for each cell. 96
6.2 Types of cells included in the Pap smear benchmark [6]. (a)-(b) Abnormal cells and (c)-(d) normal cells. 97
6.3 Results obtained in terms of HM for Fuzzy C-means classification. Notice that the HM reaches its highest value for seven features. 102

List of Tables

3.1 Advantages and limitations of state of the art methods for Pap smear cell nuclei determination. 22

3.2 Execution time for images of size 1536 × 2048. 34

3.3 Parameter Values. 34

3.4 Comparison of the proposed method with the state of the art. 39

4.1 Shape Features. 54

4.2 Texture Features. 55

4.3 Intensity Disparity Features. 55

4.4 mRMR rank of the 16 most discriminative features for the watershed, the GVF and the ACM segmentation in decreasing order with respect to their discriminative ability. The features highlighted in bold face are common for all segmentation methods. 58

4.5 Clustering Performance. 62

5.1 Results of overlapped nuclei segmentation methods in terms of Euclidean and Hausdorff Distances ($\mu \pm \sigma$) 87

5.2 Parameter values used in the experiments. 91

5.3 Execution time ($\mu \pm \sigma$). 92

6.1 Distribution of cells in the Pap-smear benchmark database [6]. 98

6.2 Features extracted from each image in the database [6]. 99

6.3 Performance of classification in terms of HM and the number of retained features. 102

Abstract in English

Marina Plissiti, E. V.
PhD, Computer Science Department, University of Ioannina, Greece. January 2012.
Tilte: Methods for Cytological Image Analysis
Thesis Supervisor: Christophoros Nikou

This thesis is focused on the development of image segmentation methods in combination with classification and clustering techniques for efficiently addressing the specific problems presented in Pap smear microscopic images. The several steps that must be followed for the the effective analysis of such images in an automated manner are described in the chapters of this thesis. As these images present great complexity and particular characteristics, the challenge for any automated methodology is to overcome the limitations of the Pap smear images. Namely, the high degree of cell overlapping, the lack of homogeneity in image intensity and the existence of many artifacts. The goal is to achieve accurate identification of the regions of interest, and as a result to obtain reliable conclusions about the contents of the Pap smear.

The processing of Pap smear images is related with several aspects of the scientific field of biomedical image processing, such as object detection, object delineation, separation of partially occluded or overlapping objects and identification of normal and abnormal figures of the object in images containing noise and artifacts. In the case of Pap smear images the objects of interest are the nuclei of the cells, as these are the structural parts of the cells which present significant changes when the cell is affected by a disease. However, the accurate detection and segmentation of the nuclei in Pap smear images is a difficult task for several reasons. First of all, in many cases the nuclei present similar characteristics with background objects. Secondly, the nuclei usually lie in areas of cell clusters, which present inhomogeneity in image intensity and the actual boundaries of each nucleus are not easily recognized. Furthermore, in cell clusters, it is common to encounter overlapping nuclei, where the borders of each nucleus are partially occluded. The correct segmentation of the nuclei is very important, at it leads to the calculation of salient features, which may contribute in the identification of abnormalities in the shape or the structure of the nucleus, in order to recognize normal or abnormal categories of the cells.

The first issue that we have successfully addressed in this thesis is the correct detection of the locations of the nuclei in images containing both isolated cells and cell clusters. The method we have developed combines global knowledge about the nucleus appearance, and local characteristics of the area and the shape of the nucleus, in order to achieve a reliable

approximation of the nuclei locations in the image. In this scope, techniques based on mathematical morphology are developed in order to detect the locations of the candidate nuclei centroids in the image. The initial rough approximations of the nuclei locations are then refined in a second step, which incorporates *a priori* knowledge about the expected shape of the nucleus and it is accomplished by the determination of the circumference of each nucleus. Finally, the elimination of the undesirable findings is achieved in two steps: the application of a distance dependent rule on the resulted centroids and the application of classification algorithms, employing features of the neighborhood of the candidate nuclei. We have examined the performance of classification techniques based on both supervised and unsupervised learning, and in all cases the effect of the refinement step improves the classification, which indicates the importance of this step.

Furthermore, based on the detection of the locations of the nuclei centroids, we have developed an automated method for the boundary determination of cells nuclei. The segmentation of the nuclei boundaries is accomplished with the application of the watershed transform in the morphological color gradient image, using the nuclei markers extracted in the detection step in order to avoid eventual oversegmentation generated by the watershed transform. For the elimination of false positive findings, features characterizing the shape, the texture and the image intensity are extracted from the candidate nuclei regions, which are used as input in a classification step, performed to determine the true nuclei. These features are tested for their discriminative ability, and a rank of the most powerful features is calculated through a feature selection scheme based on the minimum-Redundancy - Maximum-Relevance criterion. The method was evaluated on a large data set and comparisons with the segmentation results of a gradient vector flow (GVF) deformable model and a region based active contour model (ACM) are performed, which indicate that our method produces more accurate nuclei boundaries.

Concerning the separation of partially overlapped nuclei, we have developed an automated method which is based on training a physically based deformable model. More specifically, an efficient framework for the training of active shape models (ASM), based on the representation of a shape by the vibrations of a spring-mass system is employed. A deformable model whose behavior is driven by physical principles is trained on images containing single nuclei, and attributes of the shapes of the nuclei are expressed in terms of modal analysis. Based on the estimated modal distribution and driven by the image characteristics, we develop a framework, to detect and describe the unknown nuclei boundaries in images containing two overlapping nuclei. The problem of the estimation of an accurate nucleus boundary in the overlapping areas is successfully addressed with the use of appropriate weight parameters that control the contribution of the image force in the total energy of the deformable model. Comparisons with other segmentation methods, proposed especially for the separation of overlapped nuclei, indicate that our method produces more accurate nuclei boundaries that are closer to the ground truth.

Moreover, we have investigated the case of the successful classification of cells in normal and abnormal categories. Thus, a framework for the efficient classification of

cervical cells is introduced, based on features extracted exclusively from the nucleus area and ignoring the contingent cytoplasm features. This task is of high importance, since the nuclei are the only distinguishable areas in complex Pap smear images, as these images present a high degree of cell overlapping and the exact borders of the cytoplasm areas are ambiguous. We have used a database of presegmented cell images, containing both cytoplasm and nuclei features for each cell. Based on these features, we examine the ability of non-linear dimensionality reduction schemes to produce accurate representation of the features manifold, along with the definition of an efficient feature subset, and their influence on the classification performance. Two unsupervised classifiers were used and the results indicate that we can achieve high classification performance when only the nuclei features are used.

The aspects, which are discussed in this thesis, provide an integrated context for the efficient analysis of Pap smear images. The specific limitations that these images exhibit are successfully overcome and methodologies that address the restrictions in the fields of applied image analysis are described in detail. Finally, some directions for future research are also provided in this thesis.

Εκτεταμενη περιληψη στα Ελληνικα

Μαρίνα Πλησίτη του Ευαγγέλου και της Βασιλικής
Διδακτορική Διατριβή, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Ιανουάριος, 2012.
Τίτλος: Μέθοδοι Ανάλυσης Κυτταρολογικών Εικόνων
Επιβλέποντας Καθηγητής: Χριστόφορος Νίκου

Η παρούσα διατριβή εστιάζεται στην ανάπτυξη αυτόματων μεθόδων κατάτμησης κυτταρολογικών εικόνων, οι οποίες σε συνδυασμό με τεχνικές κατηγοριοποίησης και ομαδοποίησης, αντιμετωπίζουν διεξοδικά τα ειδικά προβλήματα που υπάρχουν στις μικροσκοπικές εικόνες από τεστ Παπ. Αυτές οι εικόνες παρουσιάζουν μεγάλη πολυπλοκότητα και ιδιαίτερα χαρακτηριστικά, όπως ο μεγάλος βαθμός επικάλυψης των κυττάρων, η έλλειψη ομοιογένειας στη φωτεινότητα της εικόνας και η ύπαρξη πολλών ανεπιθύμητων αντικειμένων. Έτσι, η εξαγωγή αποδεκτών και αξιόπιστων αποτελεσμάτων αποτελεί πρόκληση για κάθε αυτόματη μεθοδολογία επεξεργασίας εικόνων από τεστ Παπ. Άλλωστε, ο στόχος αυτών των μεθοδολογιών είναι η ακριβής αναγνώριση των περιοχών ενδιαφέροντος και η παραγωγή αξιόπιστων διαγνωστικών συμπερασμάτων για το εξεταζόμενο κυτταρολογικό δείγμα.

Η επεξεργασία των εικόνων από τεστ Παπ εμπεριέχει πολλά προβλήματα που η επίλυσή τους είναι αντικείμενο του γενικότερου ερευνητικού πεδίου της επεξεργασίας βιοϊατρικών εικόνων: την ανίχνευση, την κατάτμηση, τον διαχωρισμό των μερικώς επικαλυπτόμενων αντικειμένων καθώς και την αναγνώριση φυσιολογικών και μη φυσιολογικών βιολογικών σχηματισμών σε εικόνες με θόρυβο και πλασματικά ευρήματα. Στην περίπτωση των εικόνων από τεστ Παπ, τα αντικείμενα ενδιαφέροντος είναι οι πυρήνες των κυττάρων, οι οποίοι αποτελούν τα δομικά στοιχεία των κυττάρων που παρουσιάζουν σημαντικές αλλαγές όταν το κύτταρο προσβάλλεται από μια ασθένεια. Ωστόσο, η ακριβής ανίχνευση και κατάτμηση των πυρήνων στις κυτταρολογικές εικόνες από τεστ Παπ είναι μια δύσκολη διαδικασία για πολλούς λόγους. Πρώτα απ' όλα, σε αρκετές περιπτώσεις, οι πυρήνες παρουσιάζουν παρόμοια χαρακτηριστικά με αντικείμενα του υποβάθρου της εικόνας. Επίσης, τα κύτταρα συνήθως βρίσκονται σε συστάδες, οι οποίες παρουσιάζουν ανομοιογένεια στη φωτεινότητα της εικόνας και τα πραγματικά όρια των πυρήνων δεν περιγράφονται με σαφήνεια. Επιπλέον, στις συστάδες των κυττάρων, είναι συχνό φαινόμενο η επικάλυψη των πυρήνων, με συνέπεια ένα μέρος της περιοχής των πυρήνων να αποκρύπτεται. Η ακριβής κατάτμηση των πυρήνων αποτελεί ένα θέμα με ιδιαίτερη σημασία, αφού οδηγεί στον υπολογισμό σημαντικών χαρακτηριστικών που συμβάλλουν στην αναγνώριση ανωμαλιών στο σχήμα και στη δομή των πυρήνων, και συνεπώς στην κατηγοριοποίηση των κυττάρων σε φυσιολογικά και μη φυσιολογικά.

Το πρώτο θέμα με το οποίο ασχοληθήκαμε στα πλαίσια της διατριβής είναι η επιτυχής ανίχνευση των θέσεων των πυρήνων σε εικόνες που περιέχουν μεμονωμένα κύτταρα και συστάδες κυττάρων. Η μέθοδος που αναπτύξαμε συνδυάζει τη γενική γνώση που έχουμε για την αναμενόμενη μορφή του πυρήνα, και τα στοιχεία που εξάγονται από τα τοπικά χαρακτηριστικά της περιοχής και του σχήματος του πυρήνα, με σκοπό τον υπολογισμό μιας αξιόπιστης προσέγγισης των θέσεων των πυρήνων στην εικόνα. Γι αυτό το σκοπό, αναπτύχθηκαν τεχνικές που βασίζονται στη μαθηματική μορφολογία οι οποίες στοχεύουν στην ανίχνευση των πιθανών θέσεων των κέντρων των πυρήνων στην εικόνα. Η αρχική προσεγγιστική εκτίμηση των θέσεων των πυρήνων γίνεται στη συνέχεια πιο ακριβής με ένα δεύτερο βήμα εκλέπτυνσης, που ενσωματώνει τις γνωστές ιδιότητες του σχήματος του πυρήνα, και πραγματοποιείται με τον καθορισμό της περιμέτρου κάθε πυρήνα. Το τελικό στάδιο αφορά στην εξάλειψη μη επιθυμητών ευρημάτων και περιλαμβάνει δύο βήματα: την εφαρμογή ενός κανόνα απόστασης στα σημεία που ανιχνεύθηκαν και την εφαρμογή αλγορίθμων κατηγοριοποίησης, με σκοπό την κατάταξη των σημείων σε δύο κατηγορίες, την κατηγορία των πραγματικών κέντρων πυρήνων και την κατηγορία που περιλαμβάνει τα σημεία που δεν αντιστοιχούν σε πραγματικά κέντρα πυρήνων. Σε αυτό το στάδιο εκμεταλλευτήκαμε τα χαρακτηριστικά της γειτονιάς κάθε πιθανού κέντρου πυρήνα. Στα πειράματά μας, εξετάσαμε την επίδοση τεχνικών κατηγοριοποίησης που βασίζονται σε εκπαίδευση με επίβλεψη και εκπαίδευση χωρίς επίβλεψη. Σε όλες τις περιπτώσεις, το βήμα εκλέπτυνσης αποδεικνύεται ότι έχει καθοριστική σημασία, αφού βελτιώνει την απόδοση της ταξινόμησης.

Βασισμένοι στην ανίχνευση των κέντρων των πυρήνων, αναπτύξαμε μια αυτόματη μεθοδολογία για τον καθορισμό του περιγράμματος των πυρήνων των κυττάρων. Η κατάτμηση του περιγράμματος του πυρήνα πραγματοποιείται μέσω του αλγορίθμου υδροκριτών (watersheds) στην εικόνα μορφολογικής χρωματικής κλίσης, χρησιμοποιώντας ως δείκτες των θέσεων των πυρήνων τα σημεία που ανιχνεύθηκαν στο προηγούμενο βήμα, με σκοπό την αποφυγή της υπερκατάτμησης που ενδεχομένως μπορεί να προκαλέσει ο μετασχηματισμός υδροκριτών. Για την εξάλειψη των ευρημάτων που εντοπίστηκαν λανθασμένα, για κάθε πιθανή περιοχή εξάγονται χαρακτηριστικά που αφορούν το σχήμα, την υφή και τη φωτεινότητα, τα οποία στη συνέχεια χρησιμοποιούνται ως είσοδοι σε ταξινομητές, για τον καθορισμό των πραγματικών πυρήνων. Αυτά τα χαρακτηριστικά, εξετάζονται ως προς την διαχωριστική τους ικανότητα, και κατατάσσονται σε μια κλίμακα που περιέχει τα πιο ισχυρά χαρακτηριστικά, μέσω του σχήματος επιλογής χαρακτηριστικών που βασίζεται στο κριτήριο mRMR (minimum Redundancy - Maximum Relevance). Η μέθοδος εφαρμόστηκε σε ένα μεγάλο αριθμό εικόνων και εξήχθησαν συγκριτικά αποτελέσματα σε σχέση με άλλες μεθόδους κατάτμησης, που βασίζονται σε ενεργά περιγράμματα. Οι συγκρίσεις που έγιναν έδειξαν ότι τα αποτελέσματα κατάτμησης της μεθόδου μας είναι πιο αξιόπιστα και ακριβή.

'Οσον αφορά στο διαχωρισμό των μερικώς επικαλυπτόμενων πυρήνων, αναπτύξαμε μια αυτόματη μεθοδολογία η οποία βασίζεται στην εκπαίδευση ενός φυσικού παραμορφώσιμου μοντέλου. Πιο συγκεκριμένα, προτείναμε ένα αποδοτικό πλαίσιο εκπαίδευσης των ενεργών περιγραμμάτων, που βασίζεται στην αναπαράσταση του σχήματος μέσω των ταλαντώσεων ενός συστήματος “μάζες-ελατήρια”. Ένα παραμορφώσιμο μοντέλο που η συμπεριφορά του καθορίζεται από την υπέρθεση των ελεύθερων ταλαντώσεων του συστήματος “μάζες-ελατήρια”, εκπαιδεύεται σε μια σειρά από εικόνες που περιέχουν ένα μόνο πυρήνα. Με βάση την εκπαίδευση των μοντέλων και κατευθυνόμενοι από τα χαρακτηριστικά της εικόνας, αναπτύξαμε ένα πλαίσιο για την ανίχνευση και την περιγραφή του περιγράμματος άγνωστων πυρήνων, σε εικόνες που περιέχουν δύο επικαλυπτόμενους πυρήνες. Το πρόβλημα της αξιόπιστης εκτίμησης του περιγράμματος του πυρήνα στις περιοχές επικάλυψης επιλύεται με τη χρήση κατάλληλων βαρών που ελέγχουν τη συμμετοχή της δύναμης της εικόνας στη συνολική ενέργεια του παραμορφώσιμου μοντέλου. Οι συγκρίσεις με άλλες μεθόδους κατάτμησης, που έχουν προταθεί ειδικά για το διαχωρισμό των επικαλυπτόμενων πυρήνων, δείχνουν ότι η μέθοδός μας παράγει πιο ακριβή περιγράμματα πυρήνων που πλησιάζουν περισσότερο το πραγματικό όριο.

Τέλος, ερευνήσαμε την κατηγοριοποίηση των κυττάρων σε φυσιολογικά και μη φυσιολογικά, βασιζόμενοι αποκλειστικά σε χαρακτηριστικά που εξάγονται από την περιοχή του πυρήνα και αγνοώντας τα αντίστοιχα χαρακτηριστικά του κυτταροπλάσματος. Αυτή η διαδικασία έχει μεγάλη σημασία, αφού μόνο ο πυρήνας των κυττάρων μπορεί να προσδιοριστεί στις πολύπλοκες εικόνες από τεστ Παπ, λόγω της ασάφειας στον προσδιορισμό των πραγματικών ορίων του κυτταροπλάσματος εξαιτίας του μεγάλου βαθμού επικάλυψης των κυττάρων. Στα πειράματά μας έχουμε χρησιμοποιήσει μια βάση δεδομένων από εικόνες κυττάρων που η καθεμία περιέχει μόνο ένα κύτταρο και τα χαρακτηριστικά του κυτταροπλάσματος και του πυρήνα μπορούν να εξαχθούν με ακρίβεια. Με βάση αυτά τα χαρακτηριστικά, εξετάσαμε την απόδοση μη γραμμικών μεθόδων μείωσης της διάστασης των δεδομένων, για την παραγωγή μιας αξιόπιστης αναπαράστασης του πολυπτύγματος των χαρακτηριστικών καθώς και την επίδρασή τους στην επίδοση της ταξινόμησης. Τα αποτελέσματα δείχνουν ότι μπορούμε να πετύχουμε αποτελεσματική ταξινόμηση των κυττάρων βασιζόμενοι μόνο στα χαρακτηριστικά του πυρήνα.

Τα θέματα που αναλύθηκαν σε αυτή τη διατριβή παρέχουν ένα ολοκληρωμένο πλαίσιο για την αξιόπιστη ανάλυση των εικόνων από τεστ Παπ. Οι ιδιαίτεροι περιορισμοί που υπάρχουν σε αυτές τις εικόνες αντιμετωπίζονται αποτελεσματικά και οι μεθοδολογίες που επιλύουν τα προβλήματα στο πεδίο της εφαρμοσμένης ανάλυσης εικόνας περιγράφονται με λεπτομέρειες.

Chapter 1

Introduction

Cervical smear screening is the most popular method used for the detection of cervical cancer in its early stages. The most eminent screening test is the Pap smear, which is based on the staining of cervical cells, using the technique that was first introduced by George Papanicolaou [7]. This screening technique is used for over 60 years for the prevention and diagnosis of cervical cancer. Precancerous conditions, abnormal changes or infections in the endocervix and endometrium that may develop into cancer are early recognized through this routine gynaecological examination. As a result, they are early treated preventing the development of cervical cancer and the widespread use of this test in developed countries has significantly reduced the incidence and mortality of invasive cervical cancer. A regular program of Pap smear screening, with appropriate follow-up, can reduce cervical cancer incidence by up to 80% [8].

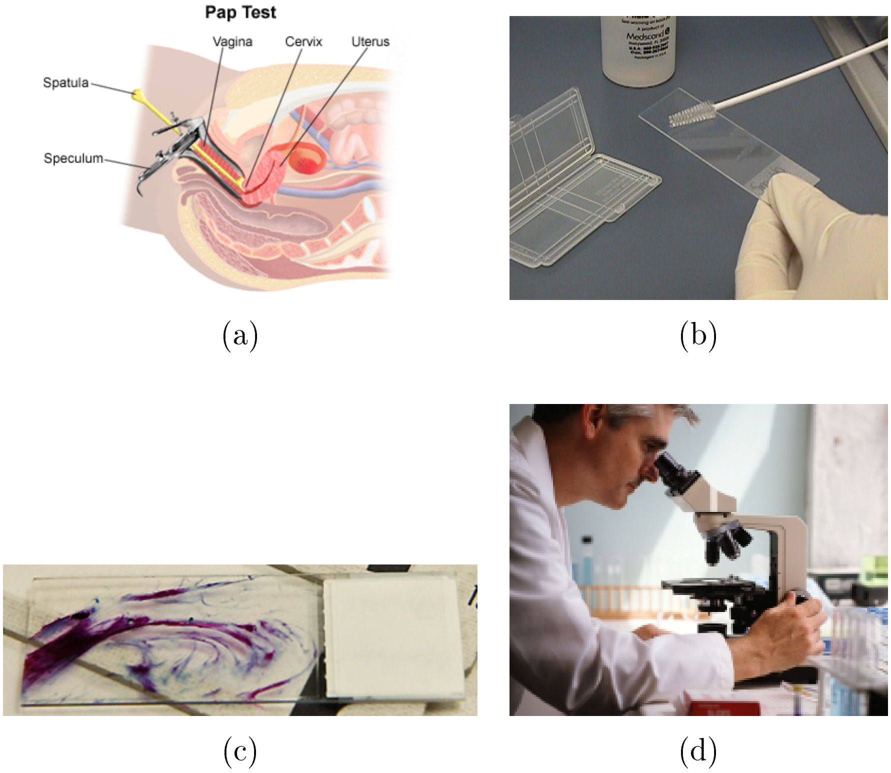
The basic steps of a conventional Pap test are depicted in (Fig. 1.1)[1]. The cervical cells are sampled and then smeared onto a glass slide and the characterization of the slide (as normal or abnormal) is accomplished through the careful microscopical examination of the slide by an expert cytologist. It must be noted that unstained cells cannot be seen with a light microscope and as a consequence they do not provide any information about their status. On the other hand, the Papanicolaou technique provides a staining procedure of cervical cells, which can be easily examined under an optical microscope. Through this procedure, tinctorial dyes and acids are selectively retained by cells. Papanicolaou chose stains that highlighted cytoplasmic keratinization in his experiments, however the diagnostic conclusions obtained through a Pap test are highly based on the nuclear features.

More specifically, normal and abnormal cells are identified by evaluating changes in the density and morphology of the structural parts of the cells, which are the nucleus

[1] The images were taken from:
http://www.reshealth.org/yourhealth/healthinfo/default.cfm?pageid=P00577,
http://redscrubs.com/2008/06/new-insight-into-cardiac-risks/,
http://www.brooksidepress.org/Products/ed2/Enhanced/Pap%20Smears/PapInterpretation.htm,
http://apps.pathology.jhu.edu/blogs/pathology/the-legacy-of-the-pap-smear-and-what-came-next

and the cytoplasm. The nucleus is the structural part of the cell that presents signifcant changes when the cell is affected by a disease. As an example, we can mention that the nucleus border abnormalities are highly correlated with the infection of cells by the Human Pappiloma Virus (HPV) [9] and the shape modifications of the nucleus are associated with the existence of Cervical Intraepithelial Neoplasia (CIN) [10]. These changes are identified through visual interpretation of the slide by an expert.

However, this manual procedure can result in an erroneously characterization of a slide as normal, while it is abnormal (False Negative). This is mainly due to the experience, stress or fatigue of the observer. Nevertheless, abnormal findings (either valid or due to technical error) usually result in considerable anxiety and in the last years many efforts have been made for computer-assisted screening and analysis of Pap smear slides. These systems attempt to provide reliable conclusions about the contents of Pap smear slides, in a fast, consistent and reproducible way.



(a)

(b)

(c)

(d)

Figure 1.1: The steps of the Pap test. (a) The sample of cells is collected from the cervix using a small cone-shaped brush, (b) the material is transferred from sampling instrument to the slide, (c) a typical Pap smear slide, (d) examination of the slide under the microscope for the detection of abnormalities.

Thus, nowadays the efficient analysis of Pap smear images is a research area of great scientific interest. Automated methodologies are proposed by several researchers, aiming at the accurate processing and classification of Pap smear images. In this scope, this thesis concerns novel techniques and sophisticated methods that have been developed and tested on a large number of Pap smear images, in order to overcome the special issues that arise from the specific characteristics that these images exhibit. A detailed description of these techniques is provided in the following chapters.

1.1 Thesis Contribution

This thesis is focused on the development of innovative and efficient image segmentation methods, which in combination with classification and clustering techniques are able to successfully address the specific problems presented in Pap smear microscopic images. These images present great complexity and particular characteristics, namely, the high degree of cell overlapping, the lack of homogeneity in image intensity and the existence of many artifacts. The analysis of these images is related with several general aspects of the scientific field of biomedical image processing, such as object detection, object delineation, separation of partially occluded or overlapping objects and identification of normal and abnormal figures of the object in images containing severe noise and artifacts. The goal is to achieve accurate identification of the regions of interest, which in the case of Pap smear images are the nuclei of the cells, and as a result to obtain reliable conclusions about the contents of the Pap smear. In this direction, automated methods for cell nuclei detection, segmentation and overlapped nuclei separation along with classification of the cells are described and evaluated in the chapters of this thesis.

In Chapter 2, a detailed description of the characteristics of the Pap smear images is provided. Furthermore, an extensive review of automated techniques for cervical cell image analysis and classification is included, in order to provide an integrated essay of the state of the art methods in the specific scientific field. Special focus has been given on two main concepts with great research interest: the cell image segmentation and the classification techniques proposed for the characterization Pap smear images. The special features and the different scientific approach for each reported method are presented in detail in both cases. Moreover, the methods that present similarities are grouped under the same heading, in order to provide a more compact description for the general techniques that have been used in Pap smear images.

In Chapter 3, a method which successfully addresses the correct detection of the locations of the nuclei in images is proposed. The method can be applied directly on images captured from an optical microscope, as it is able to detect the candidate nuclei positions in both isolated cells and cell clusters. Using the powerfull techniques that the mathematical morphology provides and exploiting all the color channels of the image, an initial rough approximation of the nuclei positions is calculated. In a second step, these approximations of the nuclei locations are refined with the determination of the circumference of each nucleus, incorporating *a priori* knowledge about the expected shape of the nucleus. The undesirable findings are eliminated with a two steps procedure: the application of a distance dependent rule on the resulted centroids and the application of classification algorithms, employing features of the neighborhood of the candidate nuclei.

We have examined the performance of classification techniques based on both supervised and unsupervised learning, and in all cases the effect of the refinement step improves the classification, which indicates the importance of this step. Furthermore, comparisons with other techniques proposed in the literature for the detection of the cells centroids indicate that the performance of our method is higher.

In chapter 4, we investigate the accurate segmentation of the cells nuclei in Pap smear images. Based on the detection of the locations of the nuclei centroids, the boundaries of each cell nucleus are determined automatically. The method is based on the marker based watershed transform in the morphological color gradient image. The nuclei markers are extracted in the detection step in order to avoid oversegmentation. In a second step, we extend the segmentation of nuclei boundaries with the determination of meaningful features of the detected areas, which contribute to the identification of the true nuclei in Pap smear images and the elimination of false positive findings. Thus, features characterizing the shape, the texture and the image intensity are extracted from the candidate nuclei regions, and they are tested for their discriminative ability through a feature selection scheme based on the minimum-Redundancy - Maximum-Relevance criterion. Comparisons of the proposed method and the results of a gradient vector flow (GVF) deformable model and a region based active contour model (ACM) indicate that the segmentation results of our method are closer to the ground truth.

In chapter 5, we propose a method for the separation of partially overlapped nuclei. The method is based on the combination of physically based deformable models, which provide a compact representation of the shape of the nucleus, and active shape models, which take advantage of the *a priori* knowledge of the expected shape. Thus, a deformable model whose behavior is driven by physical principles is trained on images containing single nuclei. The learning process deals with the attributes of the nuclei shapes, expressed in terms of modal analysis. Based on the estimated modal distribution and driven by the image characteristics, a framework for the detection and description of the unknown nuclei boundaries in images containing two overlapping nuclei is developed. The nuclei boundaries are extracted after the convergence of the deformable model. It must be noted that the contribution of the image characteristics in the energy function of the deformable model is defined by the locally adaptive image force, which is introduced in order to extract reliable nuclei boundaries in the regions of overlap. Comparisons with other segmentation methods, proposed especially for the separation of overlapped nuclei, indicate that our method produces more accurate nuclei boundaries.

In chapter 6, the classification of the cells in normal and abnormal categories is studied. As the automated detection and segmentation of the nuclei in such images has been successfully addressed in the previous chapters, we have investigated the case of the classification of cells in normal and abnormal categories using features extracted exclusively from the nucleus area and ignoring the contingent cytoplasm features. This task is of high importance, since the nuclei are the only distinguishable areas in complex Pap smear images, while the exact borders of the cytoplasm areas are ambiguous. Thus, non-linear dimensionality reduction schemes are tested for their ability to produce accurate representation of the features manifold and to define an efficient feature subset. The experiments showed that high classification performance is obtained, when only the nuclei features are used.

The several steps that must be followed for the the effective analysis of such images in an automated manner are described in the chapters mentioned above. In the last chapter, concluding remarks and some directions for future research and extension of the proposed methods are also provided.

Chapter 2

A review of cervical cell image analysis methods

Microscopic cell image analysis is one of the most significant application fields of computer vision. It concerns the analysis of images containing different samples, such as tissue (histological) or cells (cytological) images, and also images captured from different types of microscopes (optical, fluorescence etc). From the large range of these different kind of images, Pap smear images present special features, and their processing demands specialized and sophisticated techniques. In the next paragraphs, an extensive description of their characteristics and limitations is provided. Furthermore, the methods proposed in the literature are also presented, categorized by the general image processing methodology they are based on.

2.1 Pap smear images

Conventional Pap smears (Fig. 2.1) exhibit certain characteristics, such as variances in illumination and dye concentration of the cells due to the staining procedure. Also, there are numerous variables, such as air-drying, excessive blood, mucus, bacteria or inflammation which make the recognition of the suspicious cells a difficult task. In terms of screening tests, liquid-based cytology was developed as an alternative to conventional cytology, which provides more homogeneous sampling of the cervical specimen [11], and reduction in technical artifacts that limit conventional Pap test samples. However, recent systematic reviews conclude that because of the lack of well-designed comparative studies, convincing evidence to determine the superiority of the liquid based cytology for detecting high-grade lesions does not exist [12, 13, 14, 15].

The variation in cell types each Pap smear image includes is another factor of complexity of these images. There are generally three types of squamous cells seen on Pap smear images (Fig. 2.2):

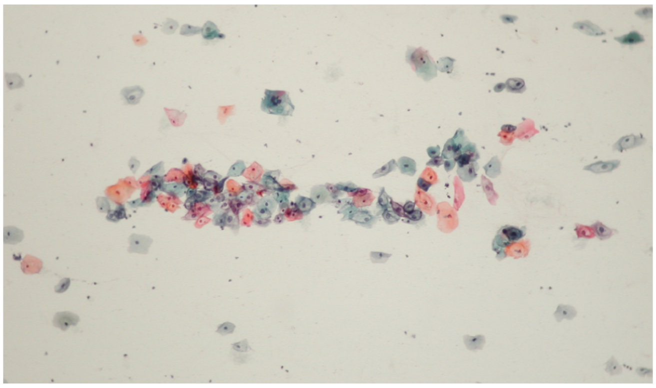


Figure 2.1: A conventional Pap smear image.

1. the superficial cells are the largest of the three types of cells and they have small pyknotic (degenerative) nuclei and cytoplasm that generally stains eosinophilic (red),

2. the intermediate squamous cells, which are similar in appearance but are slightly smaller in size and have larger, clearly structured, round nuclei with cytoplasm that usually stains basophilic (blue) and

3. the parabasal cell type that is smaller, more rounded and immature cell type.

In addition, the large number of cells (50,000-300,000 cells for an average slide), which must be reviewed by an expert, in order to characterize the smear as normal or abnormal is another important factor of difficulty. For all these reasons, significant efforts and various research approaches and algorithms have been introduced focusing on the automated analysis of cytological Pap smear images.

The accurate detection of the nuclei is crucial because the nucleus is a very important structure within the cell and it presents significant characteristics. In pathological situations, the nucleus may exhibit disproportionate enlargement, irregularity in form and outline, hyperchromasia or irregular chromatin condensation. The identification and quantification of these changes in the nucleus morphology and density contribute in the discrimination of normal and abnormal cells in Pap smear images. Thus, the challenge for any method proposed for the automated segmentation of cell nuclei is to overcome the limitations of the Pap smear images such as the high degree of cell overlapping, the lack of homogeneity in image intensity and the existence of many artifacts, in order to achieve accurate identification of the cells nuclei.

Through the segmentation of the area of the cell, several salient features can be extracted, which can contribute to the classification of the cells and therefore the characterization of the smear slide. The existence of automated classification systems would not only diminish the required time for the smear classification, but it would also avoid misclassifications due to human error. In the last years, several classification methods have been proposed for the Pap smear images, in order to identify abnormalities in the cells morphology. In general, these methods require that single cells in the slides are automatically isolated and analyzed.

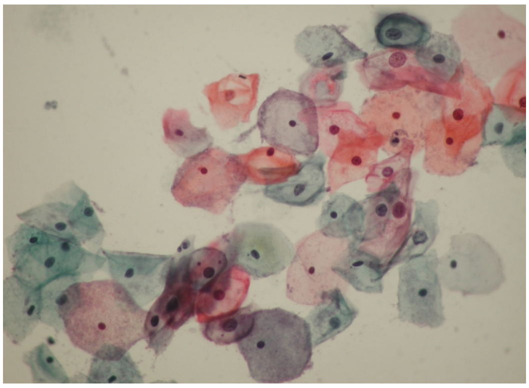


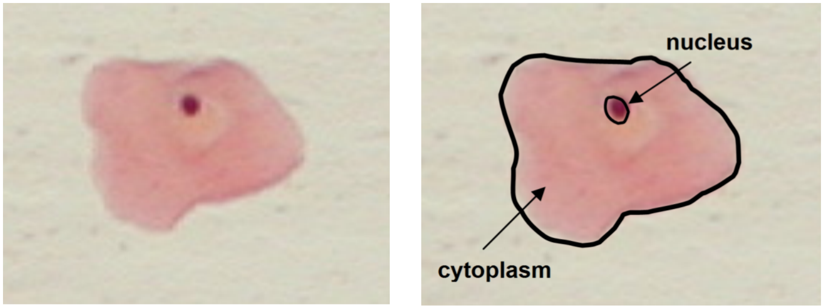
Figure 2.2: The several types of cells that are included in Pap smear images (see text for details). In this image, the high degree of cell overlapping, the inhomogeneity in cells intensity and the existence of many artifacts are clearly present.

In the next paragraphs, the segmentation and classification methods proposed for the cervical cell images are reported. The special features and the different scientific approach for each method are presented in detail in both cases. Furthermore, an attempt has been made for the methods that present similarities to be grouped under the same heading, in order to provide a more compact description for the general techniques that have been used in Pap smear images.

2.2 Segmentation of Pap smear images

The prerequisite for any further processing of Pap smear images and the derivation of conclusions for the characterization of their contents is the accurate determination of the cell nuclei area. However, the exact nuclei locations in the image and the boundaries of the nuclei areas are not clearly defined in many cases, mainly due to cell overlapping, inconsistent staining and the existence of many artifacts. For this reason, there are two open problems for every method proposed for the automated analysis of Pap smear images: the exact detection of nuclei locations and the accurate determination of nuclei boundaries. If a proposed method resolves the aforementioned restrictions, then it would be capable for direct application in images obtained from an optical microscope. However, as these images are very complicated, many segmentation methods are applied in presegmented images containing a single cell or a part of the initial microscopic image (Fig. 2.3).

The detection of the nuclei and cytoplasm boundaries in such images has been the subject of research for several scientists and many methodologies have been proposed. These methods exhibit remarkable performance in the segmentation of the structural parts of the cell. However the direct application of these methods in original Pap smear images,



(a) (b)

Figure 2.3: (a) Original cell image and (b) the structural parts of the cell; the nucleus and the cytoplasm.

which may contain a large number of cells, cell overlapping and image artifacts would probably be doubtful, without an anterior step for the determination of the locations of the cells in the image. Thus, more sophisticated approaches to the automated analysis of Pap smear images are the methods which are applied on images containing a large number of isolated cells and cell clusters, which are clearly more complicated. These methods manage to exclude the background of the image and to recognize the locations and the boundaries of the cells. It must be noted that clustering algorithms are used in this scope for the rejection of false positive nuclei positions.

In the following paragraphs, the several detection and segmentation methods are categorized in terms of the general image processing method they are based on, for the determination of the nuclei boundaries in Pap smear images. In chronological order, the first approaches were quite simple and tested on a limited data, while the most recent ones are more advanced and effective.

2.2.1 Thresholding

The first attempts to detect and segment cells in cervical microscopic images were based on image thresholding techniques exploiting the intensity histogram characteristics of the pixels [16, 17, 18]. The earliest works were proposed in the late '70s and this confirms the great scientific interest that these images attract. In general, the scope of these methods was to automatically identify threshold values in order to separate the cell from the background and the nucleus from the cytoplasm. A comparison of some threshold selection methods is included in MacAulay and Palcic [19]. In Poulsen *et al.* [20], a method for the detection of regions of interest in reduced resolution Pap smear images is proposed. The segmentation problem is transformed into an optimisation process in [21], where the determination of an optimal threshold value refers to a parametric image, which is an approximation of the initial one. A thresholding technique was also proposed in [22], for the binarization of the images and the determination of the nuclei locations, which entails in the feature extraction and the classification of the nuclei in normal or abnormal classes using a fuzzy Radial Basis Function (RBF) neural network. A multiscale local adaptive threshold method based on shape stability was proposed in [23] for the extraction of the nuclei region from the background, where the value of the threshold is obtained by preserving the expected shape of the segmented objects in the image. Finally, in [24], the detection of the nuclei is accomplished through the thresholding of the images obtained after the application of mean filters and rescaling of the gray level of the pixels. A line-scan algorithm is also proposed for the detection of abnormal nuclei.

2.2.2 Edge Detection

Several methods developed for the segmentation of the nucleus and the cytoplasm are based on edge detection, which does not require any prior knowledge of the objects of interest in the image. Thus, in [25] a cytoplasm and nucleus contour (CNC) detector is developed, which segments the structural parts of cervical cells in presegmented images. The algorithm adopts the bi-group enhancer to suppress the noises and emphasize the object contours. Then, the K-means algorithm is used to discern the cytoplasm from the background and finally, the maximal colour difference (MCD) method can precisely determine the nucleus contour.

A similar semiautomatic method was proposed by Yang-Mao *et al.* [26], where an edge enhancement nucleus and cytoplasm contour detector is presented and applied in presegmented cervical images. In a preprocessing step, three techniques, namely the trim-meaning filter, the bi-group enhancer and the mean vector difference enhancer are employed to eliminate the noise and enhance the contrast in cervical images. This results in the construction of an image having a bi-modal like intensity histogram, from which an optimal threshold can be obtained for the segmentation of the images. Finally, a relative-distance-error measure is presented to evaluate the segmentation error.

In [27], a nuclei segmentation algorithm that uses anisotropic dilation for curve closing is used. More specifically, the Canny edge detector followed by a series of morphological processes result in the detection of refined and closed edges. Then, in the edge image, a distance transform based on geodesic distances in the curved Riemmanian manifold is calculated, and a binary image is constructed after thresholding, skeletonization and morphological operations, resulting in the final segmentation of nuclei structures.

Another method based on edge detection for the segmentation of cervical images was proposed in [28]. At first, a colour depth equalization method is used in order to improve the contrast between the nucleus and the cytoplasm, and a Gaussian filter for noise removal. A two-group edge enhancement technique is then implemented based on the coarseness of each pixel. Then the Sobel operator and the non-maximum suppression are used for the extraction of the gradient image, which is binarized by setting an upper and lower edge value limit. Finally, the two longest closed curves from all the detected edges are selected to form the nucleus and the cytoplasm boundaries.

2.2.3 Mathematical morphology

The powerful techniques that the mathematical morphology provides for image segmentation are exploited extensively in some published methods aiming at the analysis of Pap smear images. In [29] a water immersion algorithm, similar in spirit with watersheds was introduced for the detection of the location of isolated cells in a microscopic image. The method is composed of three stages: the quadtree smoothing, the lowest level classification and the boundary re-estimation. The outcome of this procedure is the determination of the boundary of the isolated cells. It must be noted that this method does not provide the detection of nuclei locations in cell clusters.

Methods based on watersheds for the analysis of Pap stained images have also been proposed in the literature. In the method proposed by Jackway [30, 31], images containing one single nucleus of a Pap stained squamus epithelial cell are oversegmented with the multiscale gradient watershed transform for the extraction of texture features, in order to characterize each cell as normal or abnormal. However, the problem of the detection of the accurate nuclei boundaries is not resolved.

Several other methods based on mathematical morphology have also been proposed. In [32], sequential fuzzy greyscale morphological operations are performed for the extraction of the cell nucleus. The watershed transform has also been used in [33], where a hierarchical tree is implemented by performing multi-scale watershed segmentation on the image and then a region selection step using a measure based on spectral homogeneity and circularity identifies the regions occupied by cells in the image. In this method, a clustering step is performed, in which the cytoplasm and the nucleus regions are determined using an SVM. Finally, a method based on the watershed transform using multiscale morphological gradient and HSI colourspace was proposed in Nallaperumal *et al.*[34] , and this method was also applied in macroscopic cancer images.

2.2.4 Pixel classification

Another way of detecting the structural parts of the cells is based on pixel classification schemes. In [35], a pixel classification method based on a multifractal algorithm is developed in a first step in order to classify the pixels of the image in background, cytoplasm or nucleus. Then, an optimization step is added, by learning through genetic algorithms, and the pixels are reclassified in the above three categories. This optimization step increases the precision of the borders of the detected areas, and in the same time decreases the confusion between various classes.

As the watershed transform has been extensively used for the segmentation of the nuclei boundaries in Pap smear images, salient markers of the nuclei locations were investigated. The method proposed by Lezoray and Cardot [1] is based on pixel classification techniques for the detection of the nuclei markers, in order to avoid the oversegmentation that the watershed algorithm may produce. For this purpose, the K-means and a Bayesian classifier were used for the detection of the nuclei and other pixels class. For the later algorithm, a training set of images containing the ground truth was used for the estimation of the parameters of each Gaussian distribution.

Another approach that introduces a criterion function based on statistical structure of the objects in the image was proposed by Bak *et al.* [36], which reflects both local and global characteristics of the image. A local spatial likelihood is defined and combined with local spatial prior probabilities, producing the local spatial posterior, which is treated as a criterion function. Initially, the pixels are clustered through the K-means algorithm, and then the segmentation of the image is achieved through an iterative procedure, in which each pixel is assigned at the most probable region (nucleus, cytoplasm or background) using the predefined criterion function.

In [37], the areas of interest within the smears are detecting through fuzzy-based techniques and pixel classification. Then, the nuclei detection in images of higher magnification follows, which is also based on fuzzy rules. Moreover, in [38], a three-color based algorithm that combines colour information, experts knowledge and fuzzy systems is proposed, aiming at the improvement of the accuracy of the method proposed in [37] for the detection and the segmentation of the nuclei in Pap smear images.

Finally, a modified seed based region growing (MSBRG) algorithm for the automated segmentation of cervical cells was proposed by Mustafa *et al.* [39]. The K-means clustering algorithm classifies the pixels of the image into three categories (cytoplasm, nucleus and background) in a first step. Then from the extracted classification and using moments calculation, the location of seed pixels are determined and the MSBRG algorithm is performed. The method provides a simultaneous segmentation of all cells in the image, by identifying the pixels of the cytoplasm, the nucleus and the background.

2.2.5 Template matching

Since the nucleus in general presents an ellipse-like shape, some methods based on template matching have been proposed, for the determination of the nuclei boundaries in Pap smear images. A parametric fitting algorithm for the segmentation of cell images with application to a cervical image was firstly proposed by Wu *et al.* [40], which employs both shape and regional image information. In this work, a parametric elliptical model for the nucleus shape is introduced, and the parameters are adjusted to fit the nucleus shape by minimizing a cost function. In order to find the optimal fitting, a coarse optimization search followed by a fine optimization procedure is performed. Thus, by incorporating the shape information, the number of parameters is significantly reduced and the method results in the detection of the nucleus boundary.

Furthermore, in [41] a methodology based on a deformable template approach is developed. It consists of three steps: 1) initial estimation of the location of the cells in the image, 2) calculation of an elliptical approximation of the nucleus boundary and 3) refinement of the nucleus boundary using locally deforming models. More specifically, the detection of the location of the nuclei is obtained through a reformulation of the generalized Hough transform. Then the initial shapes of the nuclei are estimated by the determination of an ellipse. The final solution is obtained using Grenander's deformable template model [42], which converges in the correct nucleus boundary.

2.2.6 Deformable models

In general, the application of deformable models for the definition of the nuclei boundary is constrained by the requirement for the initial approximation of the model, which is essential to be close to the real boundary of the nuclei in the image. For this reason, the proposed methods based on deformable models have to solve also the problem of correct detection of the nucleus position in an image containing a large number of cells, or otherwise, their application is limited in presegmented images including only one cell.

Thus, in [43], the boundary of the nuclei is determined on presegmented images with the use of an active contour model, which is initialized through the construction of a search space, consisting of two concentric cycles. In the limits of the predefined search space, the most probable locations for the points of the nuclei boundaries are defined, following a Viterbi search-based dual active contour algorithm.

In [44], a method for the localization of cells in low resolution combined with the detection of the nucleus and cytoplasm boundaries in high resolution is proposed. The geometric active contour without re-initialization (GACWR) method is used for the localization of the cervical cells in an image of low resolution, which are then classified in free-lying cells, connected cells and irrelevant objects. After the detection of the position of each cell, the original image is partitioned into subimages containing the detected cells, and on each subimage, a binary mask is generated, in which the unwanted objects elimination and the selection of the objects of interest is achieved. In order to separate the different cells in the cell clusters, the cells are firstly modelled by a circle, which serves as the initial contour of a GACWR model. It must be noted that the GACWR is applied on an image of higher resolution, and entails in the identification of the cytoplasm. The same procedure is followed for the delineation of the nuclei boundaries, which is also initially approximated by a circle.

2.2.7 Segmentation of overlapping nuclei in cervical images

The first work concerning the recognition of overlapping cells nuclei in Pap smear images was proposed by Bengtsson *et al.* [45]. The overlap detection algorithm was based on information both from the nucleus contour and from the density profile. The method was based on the detection of significant concavities along the contour and the extraction of a set of features for the classification of the object of interest as single or overlapping. It must be noted that the nucleus contour is extracted through a thresholding procedure.

In more recent studies, the distance transform in a binary image containing the regions of the nuclei is calculated by Jung *et al.* [5] and the topographic surface generated by the distance transform is considered as a Gaussian mixture. The EM algorithm is then applied for the determination of the parameters of each nucleus cluster and the separation line of the overlapped nuclei is defined with the minimization of a criterion function. It must be noted, that in this approach, the occluded area of each nucleus is reconstructed with a constrained ellipse fitting technique.

In [4], the segmentation of clustered nuclei is treated as an optimization problem and a marker extraction scheme based on the H-minima transform is introduced to obtain the optimal segmentation result from the distance map.

2.3 Classification of Pap smear Images

The examination of the Pap smear images by expert cytologists results in the characterization of the slide, and in many cases the specific disease of the cells can be recognized. The classes that a smear slide can automatically be classified by an expert system are determined by a specialist physician. For the creation of a standardized framework for laboratory reports, the Bethesda System [46] was developed. This system was firstly introduced in 1989 and updated in 2001 as a uniform system of terminology that would provide clear guidance for clinical management. In this system, the smears are divided in two categories: normal and abnormal. The abnormal category can be further divided into four groups: atypical squamous cells (ASC), low grade squamous intraepithelial lesion (LSIL), high grade squamous intraepithelial lesion (HSIL) and squamous cell carcinoma.

Furthermore, in [6] a Pap-smear benchmark database is presented. The database consists of 917 samples distributed unevenly in 7 classes by the careful examination of skilled cyto-technicians and expert cytologists. There are two main classes that a cell may be classified: normal and abnormal. Both of these classes can be further separated into more categories. Thus, superficial squamous epithelial, intermediate squamous epithelial and columnar epithelial cells constitute the normal class. On the other hand, the characterization of the cells as mild squamous non-keratinizing dysplasia, moderate squamous non-keratinizing dysplasia, severe squamous non-keratinizing dysplasia, and squamous cell carcinoma in situ intermediate indicate the existence of abnormality in the Pap smear slide.

The characterization of the slide is mainly feasible by the evaluation of salient features of the nuclei of the cells, which, in general, is subjective to the experience of the observer. In the last years, some efforts have been made in order to obtain reproducible and objective characterization of the Pap smear images through computer vision, reducing the dependency to human experts. In this scope, several commercial interactive systems have been developed, aiming at the automated classification of the Pap smear images. These systems, such as PAPNET [47], and AutoPap [48] are based on machine learning algorithms in order to assist in cervical cytology screening.

Along with the commercial packages, the automated classification of Pap smear images has been an interesting field for the researchers and many methods have been proposed, which involve both intelligent feature extraction techniques and machine learning algorithms, in order to recognize abnormalities in these images. The basic steps of these methods are depicted in Fig. 2.4. In the following paragraphs, the features extracted from Pap smear images and the proposed techniques for their classification are described.

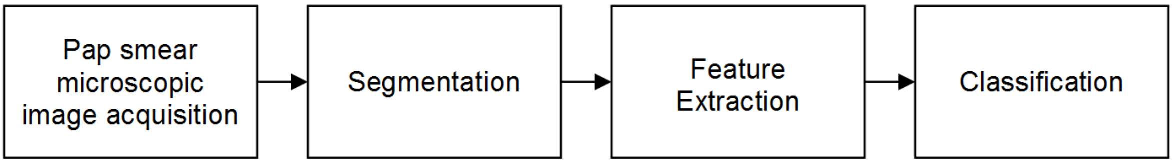


Figure 2.4: Schema of Pap smear microscopic image classification process.

2.3.1 Features of cervical cells used for classification

The published methods which deal with the classification of Pap smear images are based on the calculation of features extracted from the areas of the nucleus and the cytoplasm. It must be noted that most of these methods use presegmented images which contain only one cell, so the correct segmentation of the nucleus and the cytoplasm is feasible. In images containing cell clusters, the detection of the cytoplasm boundary is a difficult problem and until now, there is not any method in the literature that results in the automated delineation of cytoplasm in cell clusters.

The features used in these methods involve both intensity and shape characteristics of the nucleus and the cytoplasm. The most commonly used shape features are the area (size), the nucleus to cytoplasm area ratio, the shortest and longest diameter, the perimeter, the elongation and the roundness (circularity). Furthermore, the positions of the nucleus and cytoplasm centroids are used for the definition of the relative placement of the nucleus inside the cytoplasm.

Concerning the intensity characteristics, the brightness, the mean grey level and intensity disparity between the nucleus and the background are used by several techniques. In many methods, more sophisticated features are calculated in Pap smear images. Thus, in [49] a set of features extracted from the two dimensional Fourier transform of single cell images is utilized, without the need of a segmentation step, which would result in the detection of the cytoplasm and nuclei boundaries. The proposed features are based on the mean, the variance and the entropy obtained from the frequency components along the circle having as center the center of the spectrum and the frequency components along the radial line having an angle θ.

Feature vectors based on the wavelet transform were proposed in [50]. The statistical wavelet analysis and the wavelet analysis in "Brute force" approach are combined for the extraction of feature patterns. These features are then used as input in a neural network classifier for the final classification of the image in normal or abnormal categories.

Several methods of automated feature selection have been proposed for the construction of subsets of features with high discriminative ability. Thus in [51], a technique based on a genetic algorithm is proposed for the selection of the best optimal performing feature subset. The algorithmic scheme is combined with a number of nearest neighbour based classifiers, and a comparison with a tabu search based metaheuristic algorithm for feature selection is presented. Furthermore, in [52], a method based on particle swarm optimization is proposed for the feature selection problem. In this work, nearest neighbour techniques were also used for the calculation of the performance of the proposed algorithm. The results were obtained for both the two-class problem (normal or abnormal) and for the seven-class problem (as they were described in the previous section).

2.3.2 Classification techniques

The extracted features are used for training classifiers, in order to recognize and classify a certain image/cell into the corresponding class. The first attempts to classify the cells in Pap smear images were based on the Bayes rule. Thus, [53] a training set containing normal and abnormal cells is used for the calculation of the parameters of the classifier, and two decision rules are combined for the classification of each cell. In [54], the parametric Bhattacharyya distance is used for the determination of a pair of textural features and the Bayes classifier is then applied for the classification of the samples. The Bayesian classifier was also used in [49] and [55].

Another widespread technique used for the classification of Pap smear images are the artificial neural networks. In [56], a hierarchical hybrid multilayer perceptron network (H^2MLP) is illustrated for the classification of cervical images into three categories: normal, low-grade squamous intraepithelial lesion (LSIL) and high grade squamous intraepithelial lesion (HSIL). In this work, the input data are first classified as normal and abnormal by a hybrid multilayer perceptron (HMLP), and in the abnormal case, they are further classified by a second HMLP into the LSIL and HSIL categories. The performance of the H^2MLP neural network has been compared with RBF, MLP and HMLP artificial neural networks with the number of hidden nodes varying from 1 to 50. The effectiveness of the proposed diagnostic system has been demonstrated using 550 annotated Pap smear images.

The Rank M-Type Radial Basis Function (RMRBF) neural network was implemented in [57] for the classification of microscopic Pap smear images. The Median M-Type (MM) estimator was used in an RBF neural network for the estimation of the parameters of the proposed network. The results presented in terms of the sensitivity and specificity were compared with the results of simple RBF and median RBF neural network.

Several other proposed methods are based on neural networks. A feed forward neural network with a single hidden layer is used in [50] and the training method selected was backpropagation with a variable learning rate. A multilayer sigmoid neural network along with Levenberg-Marquardt backpropagation training algorithm was implemented in [58] for the classification of samples for which fuzzy based classification is unclear.

Furthermore, some classification techniques based on fuzzy logic have also been proposed. Thus, in [22], the fuzzy RBF network is applied to classify and identify the normal and abnormal nuclei. In [32], the fuzzy Adaptive Resonance Theory (ART) algorithm is used in order to classify and distinguish normal, abnormal and cancerous cells, based on the standard categories of the Bethesda System.

Support Vector Machines (SVM) were also used for the classification of Pap smear cells into normal and abnormal categories [49], [59]. Moreover, a method for cervical cancer detection using SVM based feature screening has also been proposed by Zhang and Liu [60], and it concerns the reduction of the feature space dimension in multispectral Pap smear images.

2.4 Conclusion

The task of the automated analysis and classification of Pap smear images is one of the most interesting and challenging issues in computer vision and artificial intelligence scientific fields. The identification of cervical cell nuclei areas in conventional Pap smear images is a difficult problem, as these images present great complexity and certain limitations. The accurate segmentation of the area of nucleus is a prerequisite for the derivation of diagnostic conclusions and the characterization of the contents in Pap smear images. This is feasible with the extraction of salient features which contribute in the discrimination of the cells in normal and abnormal categories by the expert classifiers.

As we can conclude, great effort has been made by several researchers in order to present effective techniques concerning both the segmentation and the classification of Pap smear images. Although the proposed techniques present high performances, the automated processing of a Pap smear slide, which would result in the documented diagnosis is not possible yet. In the future, the development of fully automated methods for the Pap smear interpretation is expected.

In the following chapters we describe our contribution in this scope. The automated methodologies that that we have proposed and developed for the efficient addressing of the nuclei detection, segmentation, classification and overlapping nuclei separation are analyzed in detail, providing qualitative and quantitative results and comparisons with some of the state of the art methods mentioned above.

Chapter 3

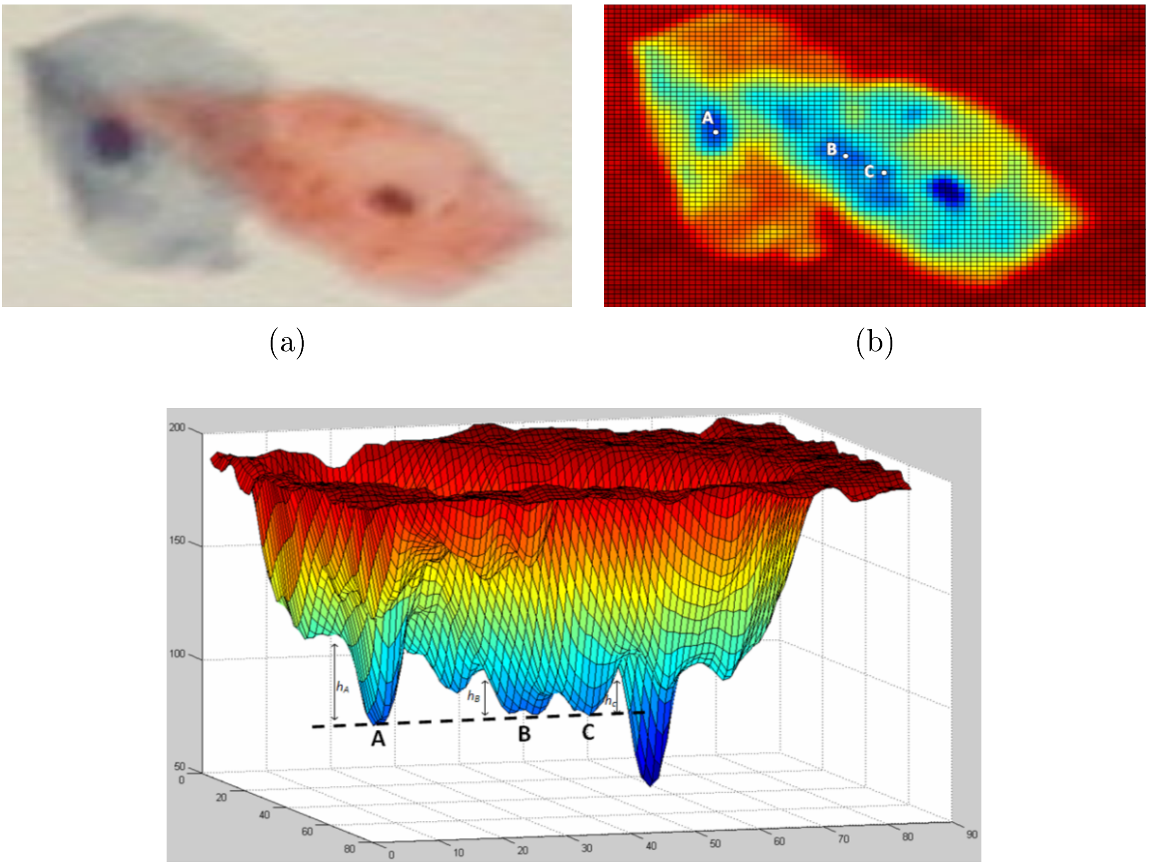
Automated Detection of Cell Nuclei using Morphological Reconstruction and Clustering

3.1 Introduction

The first attempts to detect and segment cells in cervical microscopic images were based on image thresholding techniques exploiting the intensity histogram characteristics of the pixels [16, 17, 18, 19, 21]. In addition, pixel classification was also proposed for the segmentation of cervical images [36]. Another class of methods concerns morphological watersheds for the separation of the cytoplasm and the nucleus of each cell [29, 1]. The boundaries of the structuring elements of the cells can be obtained employing methods based on active contours [43], template fitting [40, 41], genetic algorithms [35], region growing with moving K-means [61] and edge detectors [26, 28].

In Table 3.1, the methods that have appeared in the literature in the last fifteen years for the segmentation of Pap smear images are presented in chronological order. For every method, its advantages and limitations are also included. As it can be observed, many methods do not take advantage of the color information of the cervical images by converting the color image to its grayscale counterpart [40, 26, 28, 29, 35, 41, 43, 61], and therefore missing the color information. Also, the problem of overlapping cells is not considered in many methods, which are able to identify the borders of the nucleus and the cytoplasm in cervical images which contain only one cell or isolated cells [26, 28, 29, 35, 43]. Moreover, as it will be described in the next sections, most of these methods confine in presenting qualitative review for the segmentation results in several images, while their performance is usually estimated using a limited number of test images.

Considering the general methods that these approaches are based on, we can conclude that the powerful techniques that the mathematical morphology provides for the image segmentation are not efficiently exploited. Even in the case where morphological watersheds are used in [29] and [1], these methods seem to suffer from several limitations. The



(a)

(b)

(c)

Figure 3.1: (a) Initial cell image. (b) Mapping of the intensity values in the color space, where high intensity values are represented by red color and small intensity values are represented by blue color. Point A corresponds to the location of a true nucleus and points B and C correspond to areas of cell overlapping. (c) Mapping of the initial image in 3D space. The points A, B and C are lying in the same intensity level but only point A corresponds to the location of a true nucleus. As it is observed, the local depth h_A of this point is more pronounced with respect to h_B and h_C.

method proposed by Bamford *et al.* [29] was applied in gray scale images of low resolution and results in the identification of the location of isolated cells in each image. However, cell nuclei that are in cell clusters are not detected. Furthermore, the method proposed by Lezoray *et al.* [1] is based on pixel classification techniques for the detection of the nuclei markers, in order to avoid the oversegmentation that the watershed algorithm may produce. In pixels classification techniques, the choice of the number of the classes the pixels belong plays a crucial role for the final segmentation result. Pap smear images exhibit great complexity and the number of pixel classes is not obvious. The rough assumption that all the pixels of the image are distributed into two classes, such as nuclei pixels and other pixels would produce noisy results.

In this chapter, we propose a novel method for the detection of nuclei locations in conventional Pap stained cervical cell images, which may contain both isolated cells and cell clusters. The method exploits the color information as it searches for possible nuclei locations in all three channels of the color image. Furthermore, for the final determination of the true nuclei set, local color nuclei characteristics are used in a classification procedure. The method is not affected by the existence of overlapping cells and it can be applied in any cervical Pap smear image. This is confirmed by the high performance of our method in the experimental results from a large number of cervical images which contain in total 5617 cell nuclei.

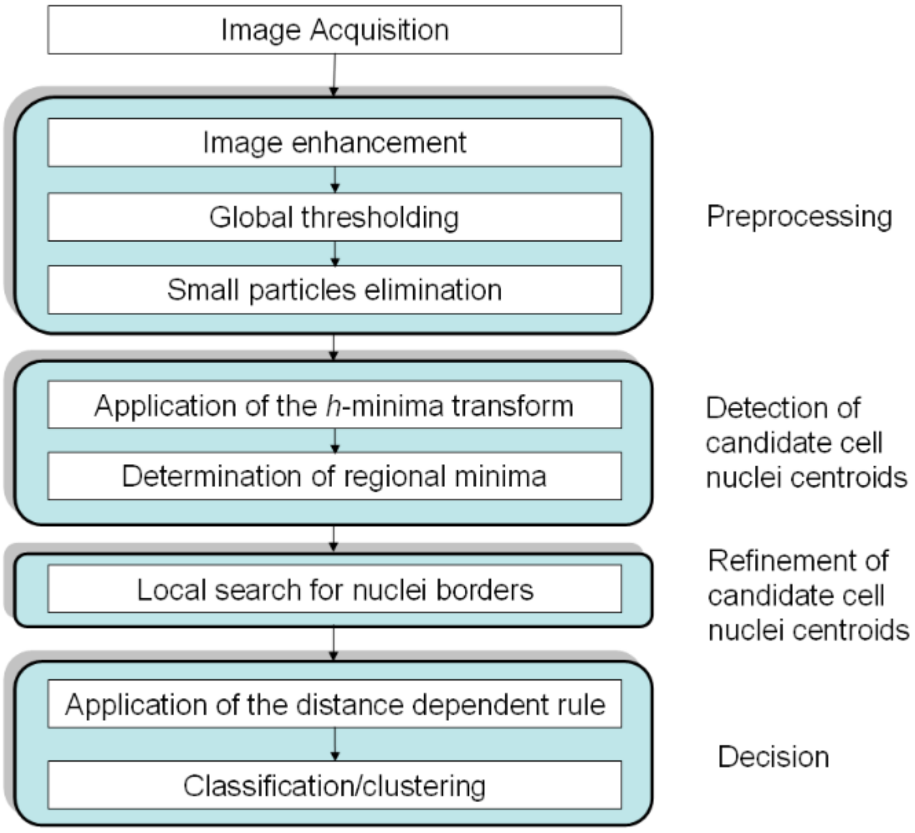


Image Acquisition
Image enhancement
Global thresholding
Small particles elimination
Preprocessing
Application of the *h*-minima transform
Determination of regional minima
Detection of candidate cell nuclei centroids
Local search for nuclei borders
Refinement of candidate cell nuclei centroids
Application of the distance dependent rule
Classification/clustering
Decision

Figure 3.2: Schema of the proposed method.

More specifically, the proposed method exploits the particular nuclei characteristics through morphological image analysis, in order to obtain automatically their locations in the image. In general, the cell nucleus is darker than the surrounding cytoplasm (Fig. 3.1(a)). However, its image intensity value exhibits extensive variation due to the staining procedure or the type of the cell and sometimes it may coincide with other areas of the image with cell overlapping (Fig. 3.1(b)). If we consider the mapping of the image in the three dimensional space (Fig. 3.1(c)), we can see that the locations of the nuclei are depicted as intensity valleys. Nevertheless, not all the intensity valleys of the same depth correspond to the location of a nucleus. As we can see in Fig. 3.1(c), the points A, B and C belong to different intensity valleys, which approximately have the same depth. However, only point A belongs to the location of a true nucleus. For the determination of the true nucleus location, the local depth of the intensity valley must be compared with the corresponding local depth of its surrounding area. This figure depicts clearly that the local depth h_A of point A has higher value than the local depths h_B and h_C of points B and C respectively. Based on this fact, we propose a method that can distinguish the true nuclei locations in Pap smear images.

Our work is summarized in Fig. 3.2. The proposed method consists of four phases: (a) the preprocessing, (b) the detection of candidate cell nuclei centroids, (c) the refinement of candidate cell nuclei centroids and (d) the decision phase which includes the determination of the final nuclei locations. In the preprocessing phase, the regions of interest in the image occupied by the cells are defined. The outcome is a binary mask which indicates the cell clusters in the initial image. In the second phase, the identification of the probable locations of the centroids of the cells nuclei takes place. A morphological-based image process is proposed for the detection of the regional minima in the image, which indicate the existence of candidate nuclei of cells. The third phase is a procedure that exploits the spatial characteristics of the nuclei and the circumference of each nucleus is approximated, which results in the refinement of the nuclei centroids locations. In the fourth phase, a decision process which extracts the true nuclei locations is proposed and it is based on two steps. In the first step, an empirical rule which depends on the distance between the centroids is applied, for the reduction of false positive occurrences. In the second step, unsupervised (fuzzy C-means [62]) and supervised (Support Vector Machines [63]) classification techniques are used in order to determine the final set of nuclei centroids. It should be noted that in the last step, we have examined the influence of the refinement of the nuclei centroids, with the construction of two data sets of patterns obtained from the initial and the refined centroids of the cells nuclei. These data sets are used as input in the classification algorithms and the results reveal that the refinement of the detected nuclei centroids is necessary for the improvement of the performance of the method.

3.2 Methodology

3.2.1 Preprocessing

In conventional Pap smear images, it is often observed that the location of the cells (isolated cells or cell clusters) is restricted in a limited space. Especially in images of low magnification (such as those obtained with 10× magnification lens), it is common that the major portion of the image is the background. However, due to a considerable amount of noise that arises from the staining process, the background is not homogenous and contains small cell particles that are not of interest. These particles must be eliminated, because they may interfere with the automated detection of cell nuclei.

In general, the preprocessing phase is necessary for the extraction of the background

Table 3.1: Advantages and limitations of state of the art methods for Pap smear cell nuclei determination.

METHOD	ADVANTAGES	LIMITATIONS
Bamford *et al.* [29]	- Simple segmentation method for the determination of the boundaries of the cells - Ensures closed boundaries	- Does not handle overlapped cells - Grayscale images - Lack of identification of the nucleus boundary
Bamford *et al.* [43]	- Ensures closed boundaries for the nucleus and the cytoplasm of isolated cells - High rate of accurate segmentation - Large number of test images	- Does not handle overlapped cells - Grayscale images - Two captures (one of low and one of high magnification) of the cell image were used
Wu l *et al.* [40]*	- Incorporates *a priori* knowledge about the shape of the cell - Investigates the case of overlapping breast cells	- Grayscale images - Many parameters to be tuned
Garrido *et al.* [41]	- A reformulated Hough transform is introduced - A deformable template model is used for the refinement of the cells boundary	- Grayscale images - The method is affected by the excess of edge points or overlapped objects in complex images
Lezoray *et al.* [1]**	- Incorporates color information on the watershed segmentation - High rate of accurate segmentation	- A training set is needed for the achievement of best results
Lassouaoui *et al.* [35]	- Introduces an optimization step based on genetic algorithms to increase the segmentation quality	- Does not handle overlapped cells - Grayscale images
Bak *et al.* [36]	- A new criterion function based on statistical structure of the object in the cell image is introduced	- Grayscale images
Mat Isa *et al.* [61]	- Region growing based technique in which the seed points locations and the threshold values are determined automatically	- Grayscale images
Yang-Mao *et al.* [26]	- A new edge enhancement nuclei and cytoplast contour detector is used - A new error measurement method is introduced	- Does not handle overlapped cells - Grayscale images
Lin *et al.* [28]	- Ensures closed boundaries	- Does not handle overlapped cells - Grayscale images

*Cervical specimen that was used for the acquisition of the test image was stained by the Crocker and Nar staining technique.

**The method was applied on images from serous cytology stained with the Pap technique.

and the definition of smooth and noise-free regions of interest, in order to reduce the searching area in the image. In our work, this step aims at accomplishing two goals: a) the definition of areas occupied by cells, and b) the rejection of areas in the image that are not of interest. In order to define the area of the cells we create a binary mask, containing the locations of the cells in the image. This mask is obtained from the initial RGB image after the application of three steps: image enhancement, global thresholding and small particles elimination in all three components of the initial color image.

In the first step, the contrast-limited adaptive histogram equalization [64] is performed individually to each color component image, which results in contrast enhancement and edge sharpening. Next, in each derived filtered image, a global threshold is obtained using the standard method proposed by Otsu [65] and with this threshold the intensity image is converted into a binary one. Finally, in the third step, the binary mask which includes the regions of interest of the image is extracted with the union of the three resulted binary images obtained from the processing of the RGB components, that is:

$$BW = BW_1 \cup BW_2 \cup BW_3, \tag{3.1}$$

where BW_1, BW_2, BW_3 are the binary masks in the red, green and blue channels of the initial image obtained with the global thresholding. The final binary image contains connected components that indicate the regions of the cells in the image. A morphological dilation is then performed in order to expand the boundaries of the region of interest and the binary mask is obtained with:

$$BW = BW \oplus X, \tag{3.2}$$

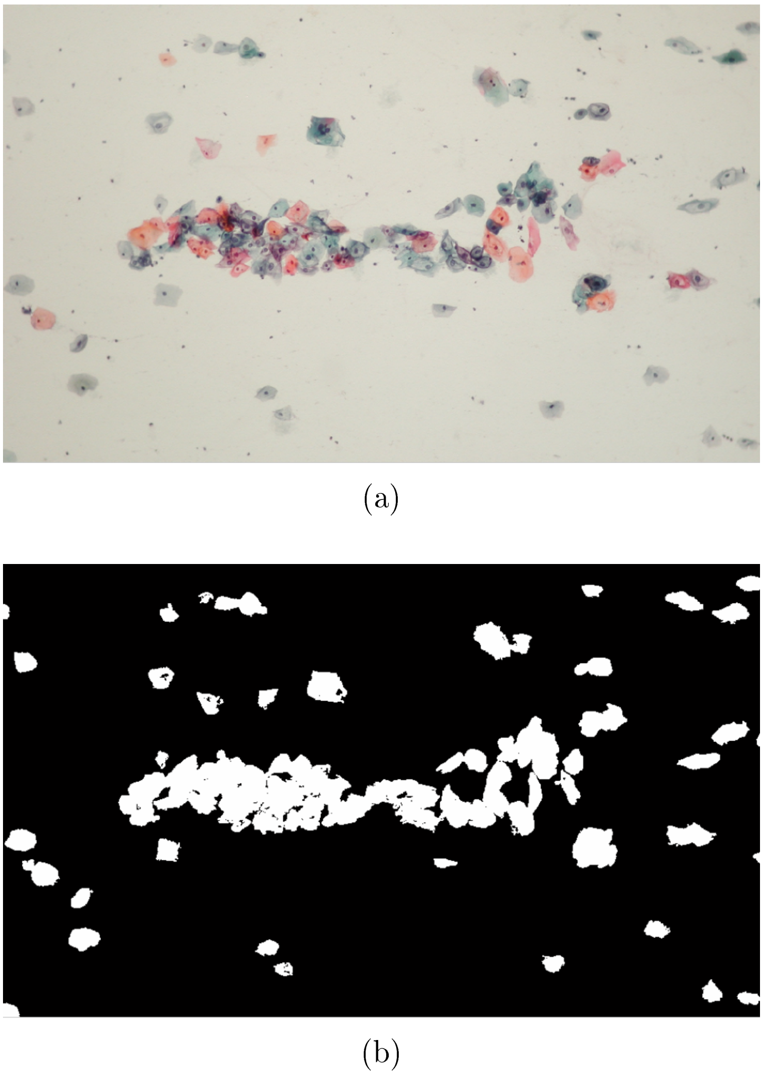
where X is a 3×3 flat structuring element.

After this operation, the connected components with an area smaller than the area of an isolated cell are undesired, because they were probably produced by the presence of image artifacts. For this reason, we remove all connected components with an area smaller than 500 pixels, which is a value smaller than the area of an isolated cell (which in general varies in the range of 900 - 7000 pixels, determined empirically after careful examination by a cytopathologist) and larger than the size of the small objects in the image. In this way, small particles are eliminated. The resulted binary image (Fig. 3.3) is used as a mask to indicate the regions that are covered by cell clusters in the initial image. Notice that it is relatively easy to eliminate the small isolated dots by the area thresholding. In these regions the detection algorithm is then applied.

3.2.2 Detection of candidate cell nuclei centroids

The areas of interest in the image obtained in the preprocessing step (Fig. 3.3(b)) contain either isolated cells or cell clusters. In the first case, the detection of cell nucleus centroid is a relatively easy procedure, as the area of the nucleus is darker than the cytoplasm. On the other hand, in cell clusters, the high degree of cell overlap and the inhomogeneities in the nuclei intensity make the detection of the nuclei a difficult task.

Our approach to this problem is based on the gray-scale morphological reconstruction [66] in combination with detection of regional minima [67] in the image. The regional minima are connected components of image pixels, whose intensity value is the same and less than the intensity value of the external boundary pixels. These minima indicate the positions of the candidate cell nuclei.



(a)

(b)

Figure 3.3: (a) The initial Pap smear image, and (b) the binary mask which is obtained after the preprocessing step.

Once we have found the regions of cell clusters, we search locally in each part of the image in order to detect the nuclei. For the definition of the search area, we compute the bounding box for each white area in the binary image and then we define the corresponding subimage in the color image. Considering that the nuclei are darker than the surrounding cytoplasm, in each subimage, we search for intensity valleys in the red, green and blue channels of the color image. These valleys consist of pixels with intensity value lower than a specific threshold, and they are bounded by pixels whose intensity value is greater than this threshold.

For the formation of homogenous minima valleys we apply the H-minima transform in the original image, which is a gray scale morphological reconstruction [68]. In this way, if the depth of each minimum is greater than or equal to a given threshold h, then the minimum is treated as a marker, otherwise it is eliminated. Thus, shorter peaks are removed, while higher peaks remain, even though they are not as significant as before.

The application of H-minima transform requires the construction of a marker image G, whose peaks determine the location of the objects of interest in the original image. A morphological reconstruction of the original image I from marker G is then performed. For the construction of the marker image G, we subtract a threshold h from every pixel of the complement I of the initial image of dimension D_I:

$$G(p) = I(p) - h, p \in D_I \tag{3.3}$$

Following the definition in [66], the grayscale reconstruction is defined with regard to the elementary geodesic dilation $\delta_I^{(1)}(G)$ of grayscale image $G \leq I$ "under" I:

$$\delta_I^{(1)}(G) = (G \oplus B) \wedge I, \tag{3.4}$$

where $(G \oplus B)$ is the dilation of G by a flat structuring element B and $\wedge$ stands for the pointwise minimum. Thus, the grayscale geodesic dilation of size $n \geq 0$ is obtained by iterating n elementary geodesic dilations:

$$\delta_I^{(n)}(G) =\delta_I^{(1)}(\delta_I^{(1)}(\delta_I^{(1)}(...\delta_I^{(1)}(G)))). \tag{3.5}$$

In this equation, the output of an elementary geodesic dilation is used as input in a new elementary geodesic dilation, and this is repeated n times. With the above definitions, the grayscale reconstruction $\rho_I(G)$ of image I from marker G is obtained by iterating grayscale geodesic dilations of G "under" I until stability is reached:

$$\rho_I(G) = \lim_{n \to +\infty} \delta_I^{(n)}(G) \tag{3.6}$$

The algorithm used for the construction of the final image is described in [66] as the fast hybrid grayscale reconstruction algorithm. The final image is the complement of the outcome image and it contains the regional minima, whose depth is less than h, suppressed (Fig. 3.4(b)).

For the determination of these regional minima, we perform the non regional maxima suppression [67] in the complement of the derived image. If we assume that $f(x)$ is the input grayscale image, F the domain of support for f and $mval$ the minimum allowed value of $f(x)$, the output image $g(x)$ is derived as follows:

1. $g \leftarrow f$
2. $\forall x \in F$
3. $if\ g(x) \neq mval$
4. $if\ \exists\, y \in Nbr(x) : g(y) > f(x)$
5. $g(z) \leftarrow mval, \forall z \in \Gamma_x\{w : g(w) = f(x)\}$,

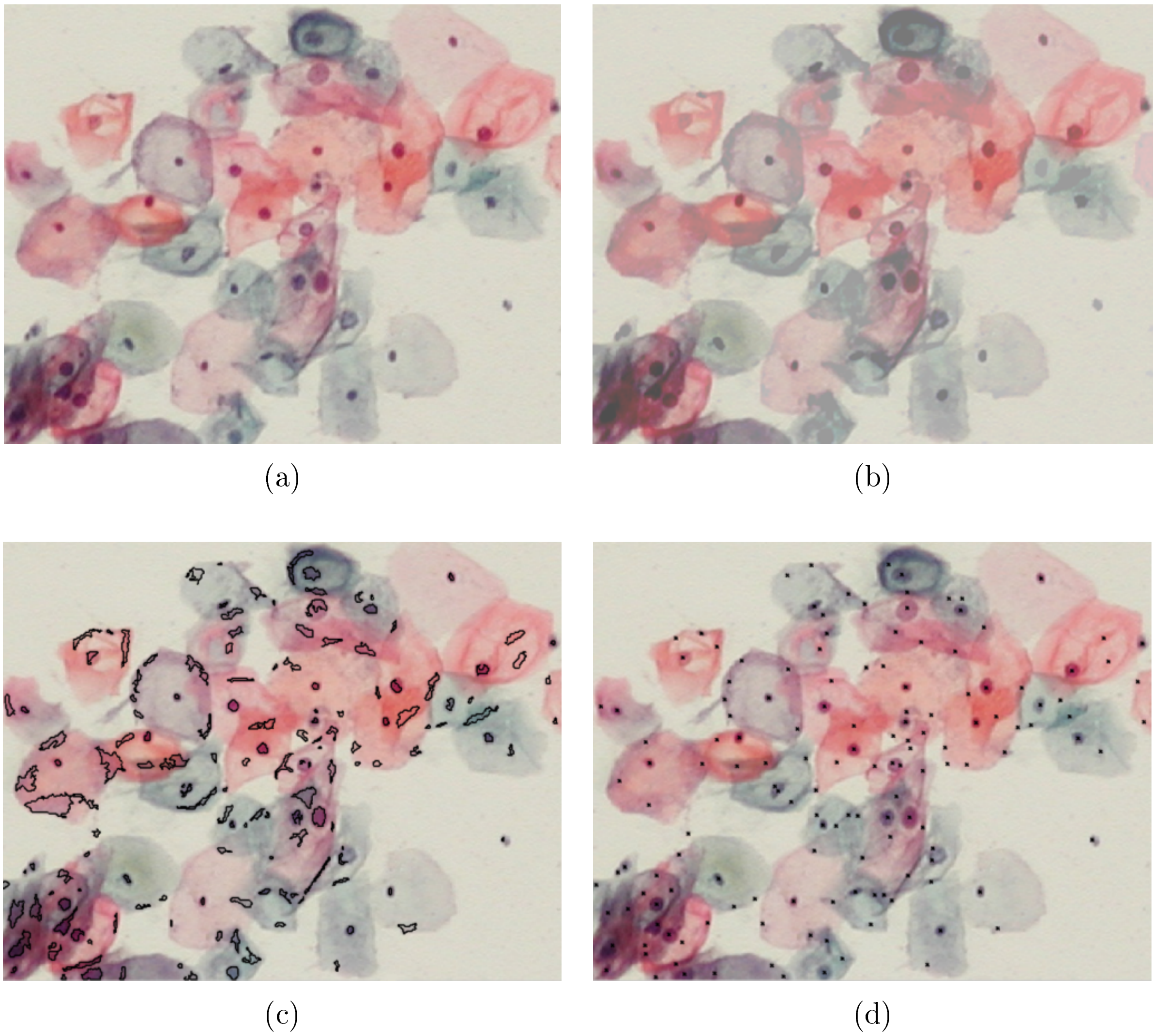

(a)
(b)
(c)
(d)

Figure 3.4: (a) Initial image of a cell cluster with overlapped cells, (b) the resulted image with the suppressed regional minima, (c) the areas of regional minima (notice that both cell nuclei and non-cell nuclei are extracted at this step) and (d) the centroids of the areas of regional minima.

where $Nbr(x)$ is the neighborhood positions associated with the image position x and Γ_x is the binary connected opening. This algorithm sets the minimum intensity value to any pixel of the image that does not belong to a regional maximum. If a pixel has a neighbor of higher intensity value, then all pixels connected to this pixel and having the same intensity are set to the minimum allowed value ($mval = 0$).

The resulted image is a binary image with the areas of intensity valleys highlighted. This procedure is applied independently in the three channels of the initial color image obtained after the preprocessing step. The areas of valleys found in the three images are joined using a logical OR operator. Following this procedure the boundaries of these valleys are calculated and the candidate nuclei are considered to be enclosed in these boundaries (Fig.3.4(c)). The location of each candidate nucleus is determined with the calculation of the centroid r_c of each detected intensity valley and it is defined as:

$$r_c = (\bar{x}, \bar{y}) = \frac{1}{N} \sum_{i=1}^{N} (x_i, y_i), \tag{3.7}$$

where N is the number of pixels consisting the boundary of the valley, and x_i, y_i are the coordinates of the pixel i of the boundary.

The list of image pixels found in this step (Fig. 3.4(d)) indicates the location of the regional minima of the image, whose depth is less than h. In general, the cell nuclei are darker than the surrounding cytoplasm and as a consequence the performance of this method is very high as it will be described in the results section. However, the centroids of these regional minima of the image do not coincide precisely with the true nuclei centroids. This happens because these minima are rough approximations of the real nuclei boundary. Moreover, as it can be seen in Fig. 3.4(d) some undesired points are detected during this step. For the detection of more accurate nuclei centroids and the rejection of regional minima centroids that do not correspond to the true nuclei locations, further processing of the image is necessary, as it is described in the following steps.

3.2.3 Refinement of candidate cell nuclei centroids

For the determination of the cell nuclei centroids we have used the global information from the cell image. However, *a priori* knowledge about the nucleus appearance in these images has not been incorporated. Most of the nuclei usually have ellipse-like boundaries, from which we can observe that the intensity of the pixels inside these boundaries are lower than those lying outside. As a result, we expect high gradient of the image across the nuclei boundaries.

Nevertheless, the value of the gradient in nucleus/cytoplasm borders varies in different parts of the image, because of the inhomogeneities in dye concentration and the variances in illumination. This is the reason why edge detectors based on the selection of a threshold in the gradient value, are inappropriate for the determination of a more precise nuclei boundary, because low thresholds would result in the detection of too many false edges, while high values would result to the loss of some true nuclei boundaries.

In this approach, we propose the use of the morphological gradient calculated with an alternative way, in order to obtain an estimation of the nuclei borders. More specifically, from the initial color image I (Fig. 3.5(a)), we construct two different images A and B. The first image A (Fig. 3.5(b)) is constructed from the original image I after the application of a grayscale erosion of the original image, that is:

$$A = I \ominus X, \tag{3.8}$$

where X is a flat disk shaped structuring element with radius 3. The use of a disk-shaped structuring element for the construction of the eroded image, pronounce the objects of the image in such a way that dark objects are enlarged radial. Thus, the nuclei become smoother and more pronounced (their area is larger and darker).

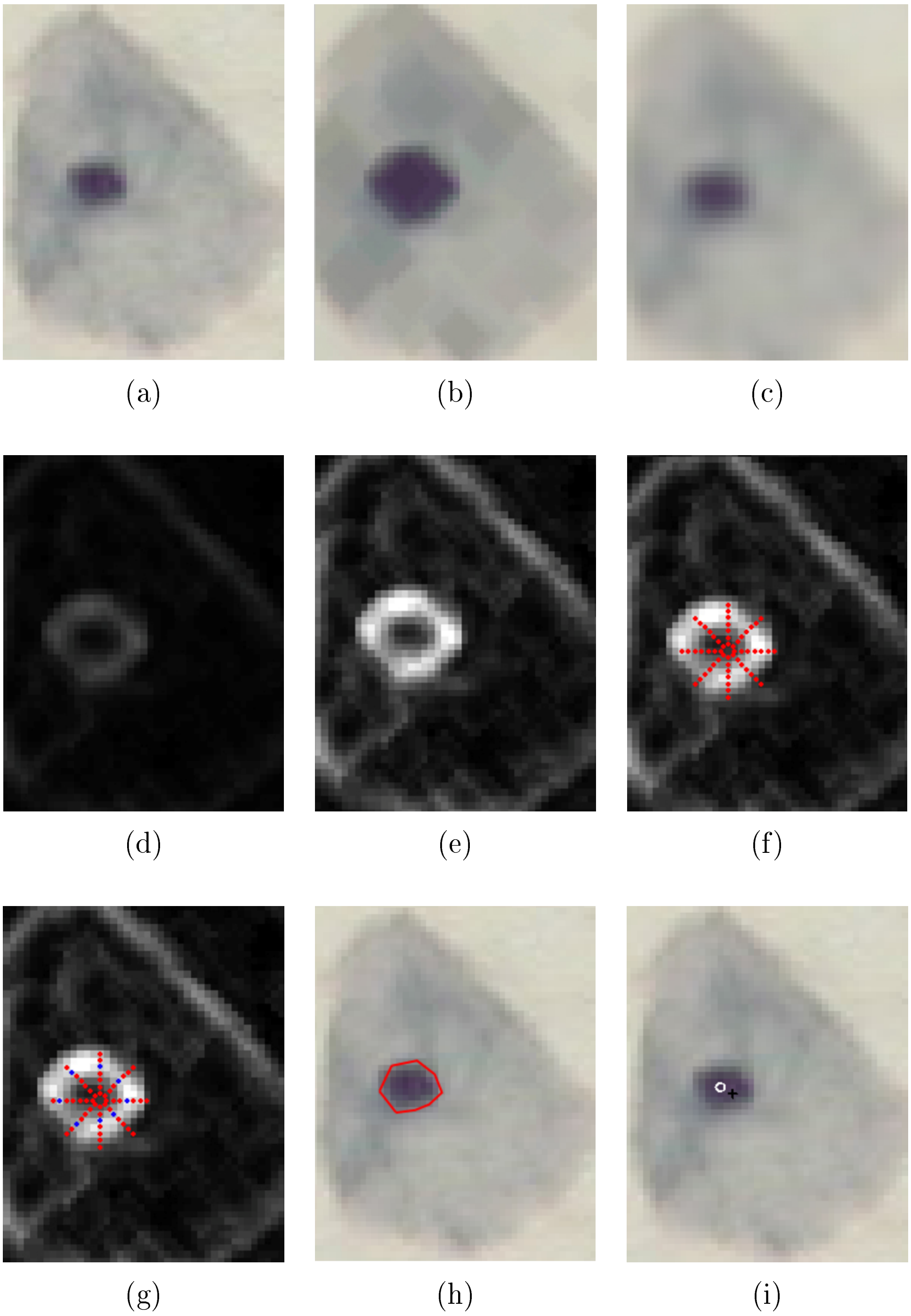


Figure 3.5: Illustration of the different steps of the refinement procedure. (a) Initial image, (b) eroded image, (c) filtered image, (d) difference of images in (b) and (c), (e) contrast enhanced image, (f) construction of the search space, (g) determination of pixels in the nucleus circumference by selecting the local maxima of the gradient amplitude, (h) the resulted nucleus contour superimposed onto initial image and (i) the initial (black cross) and the refined (white circle) centroid of the nucleus.

The image B (Fig. 3.5(c)) is the outcome of the application of a 5×5 averaging filter on the original image, given by:

$$B(x,y) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} I(i,j), \tag{3.9}$$

where $n = 5$ and $I(i,j)$ is the intensity value of the image I at the pixel (i,j) and (x,y) is the center pixel of the 5×5 region of the image. Following this procedure, noise effects and inhomogeneities in nuclei intensity are limited and a smoother image is extracted.

For the construction of the morphological gradient J of the image I, where the boundaries of the nuclei are accentuated, the subtraction of these two images is performed:

$$J(x,y) = |A(x,y) - B(x,y)|. \tag{3.10}$$

During this stage, we are interested in the determination of high intensity differences, which indicate the location of the boundaries of the nuclei, so we disregard the color information and we obtain the corresponding grayscale image (Fig. 3.5(d)). For the sharpening of nuclei borders we apply a contrast enhancement filter in the final image, which saturates 1% of data at low and 1% of data at high intensities of the original image (Fig. 3.5(e)). Finally, in the resulting gradient image, we search locally in each derived centroid for the selection of some points with high intensity values, which indicate the existence of the nucleus border.

The pixel of the initial candidate nucleus centroid is used as the starting point for the construction of a confined search space in the neighborhood of the specific regional minimum of the image (Fig. 3.5(f)). The search area in which we expect to include the boundary of each nucleus is determined using 8 radial profiles in equal arc length intervals consisted of 8 points each. After the search space is defined, in every radial profile we choose the pixel with the highest intensity (non maximum suppression, Fig. 3.5(g)). This procedure is repeated once for each candidate nucleus.

The final step is the redefinition of the nuclei centroids based on the resulted boundary pixels (Fig. 3.5(h)) using 3.7. The outcome of the entire procedure can be observed in Fig. 3.5(i). This example shows clearly that a more accurate nucleus centroid is determined.

3.2.4 Decision

The application of the method described above for the detection of nuclei centroids, produces a number of false positive occurrences, as it can be observed in Fig. 3.6 and the elimination of those points of the image that do not correspond to the true nuclei locations is necessary. This can be accomplished following a decision process based on two steps: the application of a distance dependent rule and the application of classification techniques, as it is described in the following paragraphs.

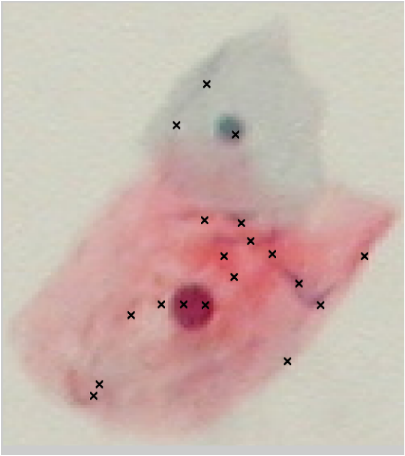


Figure 3.6: The detection of regional minima of the image results in many candidate points. This is also true for a single nucleus.

Application of the distance dependent rule

Before applying a more sophisticated classification/clustering algorithm, we have firstly applied a distance dependent rule in the set of the resulted image points. It is observed that a lot of extracted points are located in small distances between them. Even in the case of one single nucleus, the existence of more than one candidate centroid is possible and these centroids are generally spread into the nucleus circumference (Fig. 3.6). For this reason, for all the obtained centroids we apply the following rule:

```
repeat
    $\forall\, p = (x, y) \in R_c$
    if $\exists\, q = \{(x_q, y_q) | D(p, q) \leq T\}$ then
        select $r = \{p, q | min\{(I(p), I(q)\}\}$
        update $R_c$
until $no\ change\ in\ R_c$
```

where R_c is the set of all centroids, D the Euclidean distance between two points, T is the threshold on the minimum distance and $I(p)$ is the intensity of the image at point p. The threshold for the minimum distance that we use is derived from the prior knowledge we have about the true diameter of a nucleus. This is an empirical rule that is based on the fact that the points, which belong to the area of a nucleus, are usually darker than the surrounding points. By applying this rule, we have a significant reduction of the total number of the resulted centroids, while at the same time we have no loss of the true nuclei (Fig 3.7(b)).

Application of classification/clustering techniques

After the determination of the final candidate nuclei centroids, we proceed with the application of classification algorithms in order to separate the points that belong to the true nuclei and the points that belong to other regional minima, which do not indicate the existence of a nucleus in the image. Furthermore, we have examined the influence of the selection of a feature data set that is obtained using the refinement scheme of the centroids.

We have tested our method using an unsupervised and a supervised classification algorithm, namely the fuzzy C-means (FCM) [62] and the support vector machine (SVM) [63] respectively. Given the fact that the FCM algorithm does not require any training, it is applied independently in each image. However, for the application of the SVM classification algorithm a training data set is constructed, with the random selection of 34 images from the entire data set, and the remaining 4 images are used as the test set. After training, the performance of the SVM classifier is calculated using the unknown images of the test set. A brief description of FCM clustering and SVM classification algorithms are included in Appendix A. Representative results of the FCM clustering algorithm in the real image are shown in Fig. 3.7(c)-(d).

Feature vectors

The nuclei centroids obtained with the application of the proposed method in the image data set, have been detected by taking into account the general appearance and local attributes of the nuclei. For the definition of the set of nuclei patterns containing more representative features that will be used as input in the classification algorithms, we have used the intensity information which is comprised in the neighborhood of the image points found in the previous step. The dimensions of the neighborhood of the image vary in our experiments. More specifically, we have tested the performance of our method using four pattern sets of different sizes for the neighborhood, that is D1 with $3 \times 3 \times 3$ pattern size, D2 with $5 \times 5 \times 3$ pattern size, D3 with $7 \times 7 \times 3$ pattern size and D4 with $9 \times 9 \times 3$ pattern size (the third dimension corresponds to the color). Each pattern was centered at each centroid, in the initial color image. We have constructed two data sets of patterns using as the center pixel of each neighborhood the initial and the refined centroid respectively, in order to compare the performance of the method.

3.3 Results

3.3.1 Study Group

The data set that is used in this work is composed by conventional Pap stained cervical cell images, acquired through a CCD camera (Olympus DP71) adapted to an optical microscope (Olympus BX51). We have used a 10× magnification lens and the acquired images with size 1536 × 2048 were stored in JPEG format. We have collected 38 images from 15 Pap smear slides and the total number of cell nuclei in the images is 5617. In order to obtain the ground truth, the nuclei locations were manually identified by two expert cytopathologists.

3.3.2 Numerical evaluation

Our method is fully automated and the final detected nuclei are obtained even in areas with high degree of cell overlapping such as the cell clusters in the image (Fig. 3.4), without any user interference. For the evaluation of the performance of the method we have to examine the performance of the different steps of the method, starting from the preprocessing step until the application of the classification algorithms, in which we have tested several different parameters. Furthermore, as a measure of the computational efficiency of the segmentation method, we present in Table 3.2 the processing times of the individual steps of the method developed in Matlab using a Pentium 2.0 GHz with 3GB RAM.

The preprocessing is a fast procedure which results in the determination of the region of interest in the image, since it excludes all the background and leaves, for further processing, the parts of the image containing isolated cells or cell clusters. It misses 9 cell nuclei in all images and it produces a reduction of true positives cell nuclei of 0.16% of the total initial number of nuclei. For this reason, the number of cell nuclei which must be detected in the following steps is decreased to 5608.

The detection step of the cell nuclei centroids has shown that the detected points have successfully identified the location of most of the nuclei in the image, as it is confirmed by the expert observer. During this step, 42 true nuclei centroids are missed and the true nuclei detection rate is 99.25%.

The application of the distance dependent rule on the refined nuclei centroids yields in the reduction of false positive findings at the rate of 14.13% while we have no loss of true nuclei centroids. This rate could be higher if we select a distance threshold higher than 8 pixels. However, with a selection of a higher value for this threshold, true nuclei centroids are missed, as it can be observed in Fig. 3.10.

For the application of the classification algorithms, we have used two data sets, as it is already described. In FCM algorithm we have used the Euclidean and the diagonal norm as the distance-dependent metric. The Euclidean norm is defined as:

$$D_{ij}^{Euc} = \sqrt{\sum_{j=1}^{c}(x_{ik} - v_{jk})^2},\ 1 \le k \le p,\ 1 \le i \le N, \tag{3.11}$$

where N is the number of the x_k unlabeled column vectors in R^p, p is the number of features in each vector x_k, c is the number of different clusters and $\{v_i\}_{i=1}^{c}$ are the prototypes of the clusters. Respectively, the diagonal norm is defined as:

$$D_{ij}^{Diag} = \sqrt{(x_i - v_j)^T A_D (x_i - v_j)},\ 1 \le i \le N,\ 1 \le j \le c, \tag{3.12}$$

where $A_D = diag(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, ..., \frac{1}{\sigma_p^2})$ and σ_i is the standard deviation of each measured characteristic. The results of the classification are presented in Fig. 3.8.

Furthermore, we have trained the SVM classifier with a training set of 34 randomly selected images and the performance of the SVM classification is calculated using the 4 remaining unknown images from our dataset. This experiment was repeated 20 times, and the average performance was calculated. The results are summarized in Fig. 3.9.

For the comparison of the results we have calculated the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) findings of each classification technique. Then, two widely used statistical measures for the performance of the classification are calculated:

1. The *Sensitivity*, which measures the proportion of actual nuclei which are correctly identified as such and it is defined as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.13}$$

2. The *Specificity* which measures the proportion of candidate centroids that are not nuclei and are correctly characterized as such by the classification techniques, and it is defined as:

$$Specificity = \frac{TN}{TN + FP} \tag{3.14}$$

These measures were selected in order to evaluate the ability of the method to recognize the true nuclei and the true artifacts from the total number of findings in the image.

As it is depicted in Fig. 3.8 and Fig. 3.9, the FCM algorithm has higher *Sensitivity* than the SVM, which means that fewer true nuclei are missed by the algorithm classification. However, the *Specificity* of FCM is low relatively to SVM, which means that FCM includes a lot of false positive centroids in the final set of the points characterized as nuclei centroids by the algorithm. On the contrary, the *Sensitivity* of the SVM classification is relatively low, namely it misses more true nuclei centroids during the classification. Nevertheless, it presents high *Specificity* rate which means that in the final set of points characterized as nuclei, the false positive occurrences are limited.

An important fact that must be noted is that, as we can see in Fig. 3.8 and in Fig. 3.9, in both unsupervised (FCM) and supervised (SVM) classification techniques, the use of the refined centroid data set leads to a better classification performance. This indicates that with the refinement step, the prior knowledge of the nucleus shape that is incorporated leads to more accurate localization of the nucleus centroid. The contribution of this step is crucial because the refined centroids are closer to the true nuclei centroids and the datasets of patterns that are constructed contain more representative features of the nuclei centroids neighborhood. This results in the improvement of the discrimination ability of the classification techniques, since the patterns of the true nuclei class are more compact as they contain only pixels from the nucleus area and they do not include pixels from the cytoplasm area.

Table 3.2: Execution time for images of size 1536 × 2048.

Step of the method	Processing time in sec. (mean ± std)
Preprocessing	3.53 ± 0.22
Detection of candidate cell nuclei centroids	72.45 ± 39.55
Distance dependent rule	7.06 ± 13.00

3.4 Discussion

3.4.1 Evaluation of the proposed method

The proposed method is fully automated and its application was performed without any observer interference. The method consists of four individual stages: the preprocessing, the detection of candidate nuclei, the centroids location refinement and the decision step, which results in the final determination of the cell nuclei locations. As it is verified by the results, the method is suitable for the detection of cell nuclei in Pap smear images, even when cell overlapping is present.

Table 3.3: Parameter Values.

Step of the method	Parameter	Value
Preprocessing	Area threshold for the elimination of small particles	500
Detection of candidate cell nuclei centroids	Intensity depth threshold (h)	15
	Minimum allowed Image value (mval)	0
Refinement of candidate nuclei centroids	Number of radial profiles	8
	Length of radial profiles	8
Distance Dependent Rule	Minimum Distance threshold (T)	8
FCM	Weighting component (m)	2
SVM	Linear Kernel: C	0.001
	RBF Kernel: C	1
	RBF Kernel: γ	1

Concerning the parameters involved in the algorithm (Table 3.3), from the entire image data set (38 images) we have used 19 randomly selected images which contain in total 3616 cells nuclei, in order to estimate the values of these parameters, which correspond to the prior knowledge of the cells nuclei. After careful examination of these images from an expert cytopathologist, the area threshold of 500 pixels was considered to be sufficient for the small particles elimination in the image. Thus, connected components of area smaller than 500 pixels are characterized as objects of no interest, while the isolated cells in the image are preserved. Nevertheless, the loss at this step is due to the fact that some of the cell cytoplasms are faintly stained, and they are not distinguishable from the background. As a consequence, the nucleus is considered as an isolated object, and with the application of this step, it is removed.

The selection of the threshold of the depth of the intensity valleys influences the number of true positives and false positives. For the choice of the threshold we have performed several tests using different values, and we chose the threshold value 15, which produce the minimum loss of true positives centroids. As it is depicted in Fig. 3.11, with this threshold value we obtain the maximum number of true nuclei centroids detected, while the number of false positive centroids is kept at a low rate.

The outcome of this step is the detection of image regional minima, which may indicate the locations of cell nuclei. As it is observed in many cases, the projection of these centroids in the images does not coincide with the true nuclei centroids (Fig. 3.5(i)). This occurs because the determination of the initial centroid is a coarse approximation of the true nuclei centroid. With the exploitation of the *a priori* knowledge of the nucleus shape characteristics in the refinement step, we obtain more accurate nuclei centroids in the image. We search for 8 points in the nucleus circumference and for the construction of the search space we use radial profiles of 8 pixel length, as this was estimated to be the average size of the nuclei radius by the expert observer. By these means, we obtain a smooth nucleus boundary, and based on this, we calculate a new refined nucleus centroid.

The result of this step is the extraction of nuclei markers in the image. However, in some areas of the image it is observed that two or more markers are lying on the same nucleus (Fig. 3.6). This is a consequence of the inhomogeneity in the nucleus dye concentration. The elimination of these additional markers is necessary and may be achieved by the application of the distance dependent rule. If we omit this step, then there will be some centroids which belong to the same nucleus and they will introduce interference in the clustering step. For instance, if they are classified in the same class (e.g. the nuclei class) we will not be able to compute the number of true detected nuclei, since one single detected nucleus will be counted twice. On the other hand, if they are assigned to different classes (one centroid in the nuclei class and the other one in "other findings" class) then one centroid will be counted as true positive and the other one as false negative. This would be wrong, since they are both lying in the area of the same nucleus. For this reason, the distance dependent rule is necessary, in order to overcome such configurations. We calculate the Euclidean distance of each marker from its neighboring markers and the threshold we use is 8 pixels, as it approximates the average nucleus radius in our images. In this step, we achieve to maintain all the true nuclei locations. After the determination of these values, the method was applied in all 38 images of our data set and the obtained results indicate that the selection of these values corresponds to a reliable estimation of the shape and the intensity characteristics of the cells nuclei.

In Table 3.2 the processing times of the individual steps of the method are provided. As we can see, the execution time varies significantly in the detection of regional minima and in the application of the distance dependent rule. This is a consequence of several factors. First of all, the number of the real cell nuclei that each image includes plays a crucial role in the execution time of the detection of regional minima step. As the number of the nuclei in our image data set varies from 26 to 522 nuclei, we expect high variation in the execution time. Furthermore, the result of the preprocessing step and the proportion of the image that is identified as background is another factor that influences the execution time. In an image with smooth background and well stained cells, the region of the image that is excluded as background is large and the cells clusters were successfully recognized. However, in an image with artifacts, severe noise and variation in cell staining, the preprocessing step results in the selection of some regions of the image, that do not correspond to the true location of the cell clusters. Even that in those areas no cells are present, the regional minima detection step is also performed in those areas, and this demands additional execution time. This also results in the detection of false positive findings, which affects the execution time of the distance dependent rule step, because more candidate points are processed. Finally, the variation of the execution time of these steps is affected by the presence of outlying images which exhibit high difference from the mean execution time and although are a few, their influence in the total variation is significant. All of the above factors lead to a distribution of execution times having a central concentrated blob around its mean value and several sparse cases with very high execution times, thus leading to an increased standard deviation of the run time in the set of the experiments.

The final step of the proposed method is the application of the clustering/classification techniques (FCM or SVM) for the extraction of the final set of true nuclei centroids. As it is already mentioned, in FCM we used two clusters and the weighting exponent m is set to 2. The values of the parameters in SVM are $C = 0.001$ for the linear kernel type, and $C = 1$ and $\gamma = 1$ the RBF kernel type. The training of the SVM classifier leads to the selection of some tens of support vectors, depending on the type of the kernel, the data set and the dimensions of the patterns that we use.

Considering the classification performances, the selection of one of the above classification techniques (FCM or SVM) depends on the purpose of the detection of nuclei in a specific Pap smear image. For instance, if the purpose is to find abnormal or malignant cell nuclei, the FCM is preferable, as it produces lower loss of true nuclei and the probability of a missed abnormal nucleus is reduced. On the other hand, if the purpose is to detect cells nuclei in order to calculate, for example, morphological characteristics, a pure set of true nuclei would be desirable and the SVM classification technique is suitable, as it reduces the false positive occurrences in the final set. However, since the performance of SVM depends on the selected values of the parameters, its use becomes more demanding, especially when a limited number of images exist. On the other hand, the application of the FCM algorithm can be performed directly in one single image. As a result and in combination with the high performance it presents, the FCM algorithm is preferable for the classification step of our method.

3.4.2 Comparison of the proposed method with other methods

We have also compared our method with the detection methods proposed by the state of the art technique of Lezoray *et al.* [1] which is based on pixel classification schemes. More specifically, the k-means clustering algorithm and a Bayesian pixel classification scheme were applied to our image data set following the principles in [1]. It must be noted that these schemes were performed in the preprocessed images, with the background removed, and they classify each pixel as “nuclei” pixel or “cytoplasm” pixel. The application of the k-means classification algorithm does not need any training and it is applied directly in every image of our data set. However, in Bayesian pixel classification, some parameters must be determined. Thus, the *a priori* probabilities of each class are considered as equal and all the parameters of the mixture of Gaussian distributions were calculated on a training set of color vectors, which were derived from randomly selected images of our data set (50% of the images), as it is also proposed in [1]. More specifically, in each preprocessed image, the pixels of the nuclei locations and the pixels of the cytoplasm were used for the calculation of the mean and the covariance matrix of each Gaussian distribution. Then, the Bayesian classifier was used for the classification of the pixels in the remaining images. This experiment was repeated five times, each time with a different (randomly selected) training set.

The outcome of both the k-means algorithm and the Bayesian pixel classification schemes was the determination of some regions in the images that would probably be occupied by the nuclei. These regions are considered as connected components and they are compared with the connected components extracted from the detection of regional minima of our method, in terms of how many true nuclei centroids were recognized. Thus, the number of the true nuclei centroids that each connected component contains was calculated, in order to compare the performance of each method. The expected results would be the detection of one nucleus per connected component, since the existence of more than one nucleus centroid in one connected component does not provide the correct localization of the nuclei, as it leads to erroneous estimation of the nucleus centroid. Thus, the desirable performance of each method is a high number of connected components that contain only one nucleus. In Fig. 3.12, we can observe the average number of the detected connected components, over the test sets of images, which were recognized by the compared methods. As we can see, our method is superior compared to the pixel classification schemes, since it produces more single connected components which contain only one nucleus, while the other methods may result in the detection of connected components that contain even 10 true nuclei centroids. This is a major drawback of the standard pixels classification schemes, as they can not detect accurately a big number of the true nuclei centroids of our image data set. Let us also notice that the vertical axis in Fig. 3.12 has a logarithmic scale making the differences in performance more pronounced.

Beyond the comparison of our method with pixel classification schemes, Table 3.4 shows a comparison of the proposed method and other methods appeared in the literature for the segmentation of cell nuclei in cervical images. In general, it is difficult to compare the methods directly since many of the methods do not include quantitative results and the performance criteria that they use vary extensively. Furthermore, some data parameters are not clearly defined, such as the number of different smear slides, the number of images captured from the optical microscope, the size of the test images and the number of cells that the images contain. These parameters are important for the evaluation of the general behavior of each method, when it is applied in a large number of different data resources.

From Table 3.4 we can assert that the performance of our method is superior for several reasons. First, the data set that was used includes images captured from different Pap smear slides taken from 15 individual patients, which evince that the data set contains a big variety of different cells and the obtained results describe more precisely the general behavior of the method and the expected performance in a new image. Also, the proposed method can be applied in images captured directly from an optical microscope with size 1536×2048 and is able to successfully recognize the cells nuclei, even in cases where cell overlapping is present. Thus, there is no need to presegment the individual cells of the image in order to find the nuclei locations. Moreover, the average number of cells nuclei in these images is 148, and they are clearly more complicated than those images that contain only isolated cells, such as in methods [26, 28, 29, 35, 43].

In terms of the general image processing approach, the method exploits the color information of the image, in contrary to the techniques in [40, 26, 28, 29, 35, 41, 43, 61]. This is advantageous, since the staining process of the smear has different effects in the three color components of the image and some nuclei are more distinguishable in a single color channel. The use of three different thresholds (one for each color channel) in the Otsu's method in the preprocessing step is more effective than the use of one single threshold in the grayscale image. Furthermore, the detection of the intensity valleys in the three channels of a color image and the merge of the detected regions in a final image results in the determination of more true nuclei locations, rather than the detection of the intensity valleys in the grayscale image. To underpin these observations, in Fig. 3.13 we depict the results of the steps of the proposed method, when it is performed in a color and the corresponding grayscale image. As we can see, both the preprocessing step and the regional minima step fail to recognize the same number of the true nuclei in the grayscale image, as it is successfully achieved in the case of the application of the method in the color image. The individual processing of each color component and the combination of the results in all three components leads in no loss of information and in the determination of more accurate nuclei centroids.

3.5 Conclusion

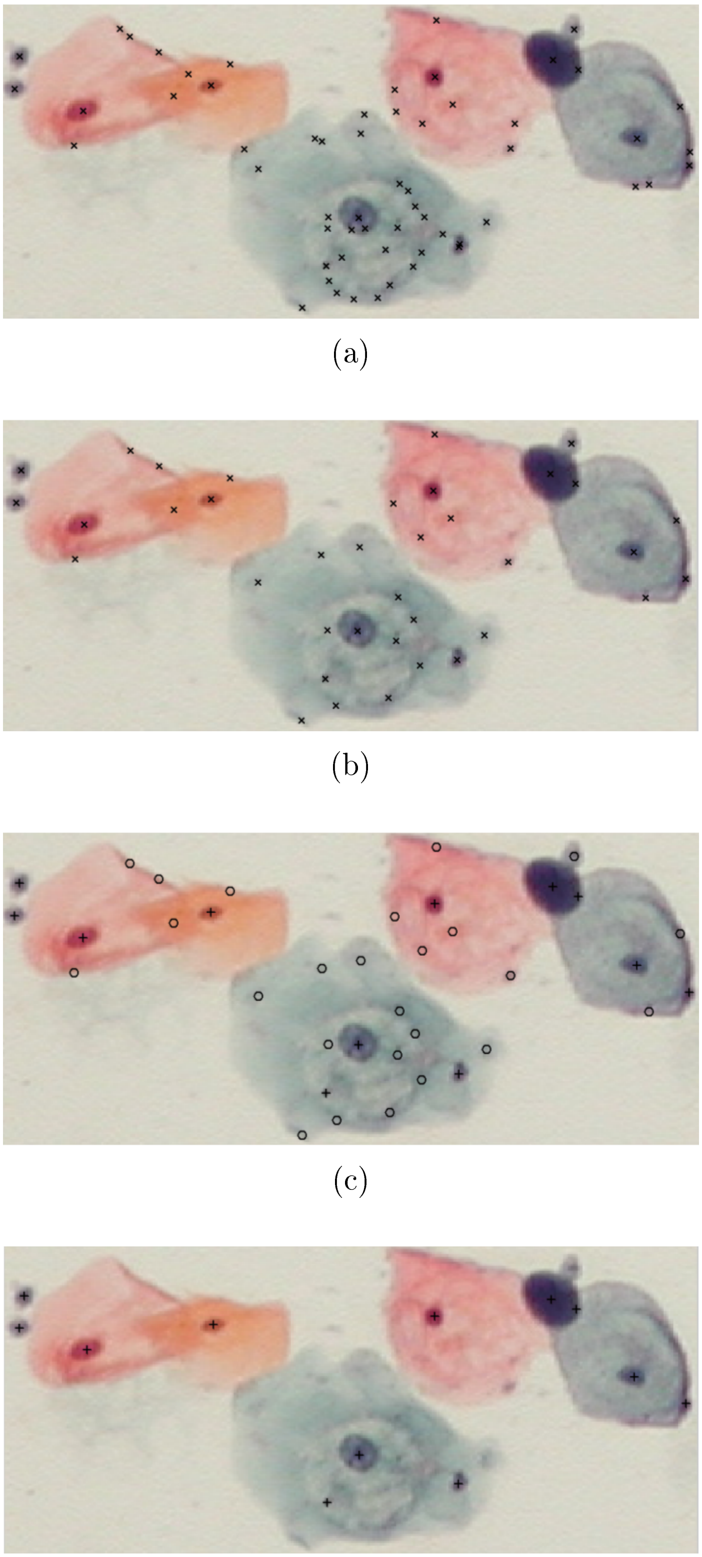
The task of identifying the cell nuclei in conventional Pap smear images is a challenging issue, especially when it must be automated. From this point of view we propose a robust and accurate method. Given the fact that our image data set derive from different Pap smear slides, we expect our method to give acceptable results and to present high performance, when it is applied in a new Pap smear image. Moreover, the proposed method can be used as the basis for further processing of cell images, which is a non trivial and time consuming procedure for expert observers. The major advantage of the proposed methodology is that it is fully automated and it is suitable for cell images with high degree of cell overlapping, as it can detect not only the nuclei of isolated cells but

Table 3.4: Comparison of the proposed method with the state of the art.

Method	#Slides	#Images	Size	Cells	Performance Criteria	Quantitative Results
Bamford *et al.* [43]	-	-	128×128	20130*	Visual inspection	99.64% correctly segmented cells
Wu *et al.* [40]	1	1	80×100	1	Comparison with K-means and Bayes classifier in a synthetic image	Misclassification rate lower than 5%
Garrido *et al.* [41]	-	3	-	-	Visual inspection	Lack of quantitative results
Lezoray *et al.*[1]	-	10	-	209	Vinet measure Number of segmented regions	Mean Vinet measure of 2.24 for RGB and 3.41 for HSL Mean difference of the segmented regions and manually segmented regions of 2.87% for RGB and 0.47% for HSL
Lassouaoui *et al.* [35]	-	2	256×256	-	Visual inspection	Lack of quantitative results
Bak *et al.* [36]	-	2	-	-	Visual inspection	Lack of quantitative results
Mat Isa *et al.* [61]	-	3	-	-	Visual inspection	Lack of quantitative results
Yang-Mao *et al.* [26]	-	-	64×64	124*	Misclassification error, edge mismatch, relative foreground area error, modified Hausdorff distance, region nonuniformity, relative distance error	Average segmentation error of 0.1145
Lin *et al.* [28]	-	10	-	10*	Misclassification error, relative foreground area error, modified Hausdorff distance	Average segmentation error of 0.1323
This work	15	38	1536×2048	5617	*Sensitivity (Se)* *Specificity (Sp)*	Indicative mean values are *Se* = 90.57%, *Sp* = 75.28% for FCM and *Se* = 69.86%, *Sp* = 92.02% for SVM

*These numbers correspond to presegmented individual cells (one cell/test image)

also the nuclei in cell clusters with high sensitivity.



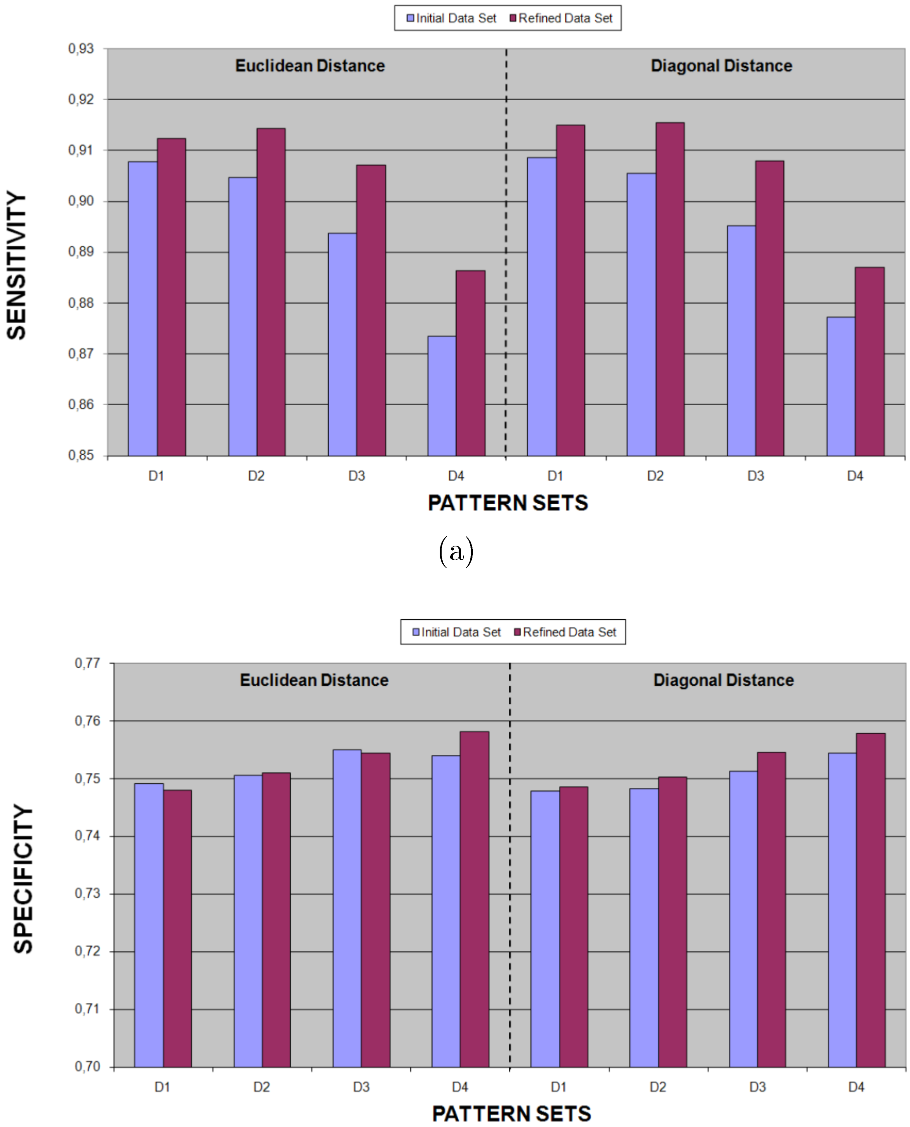
(a)

(b)

(c)

(d)

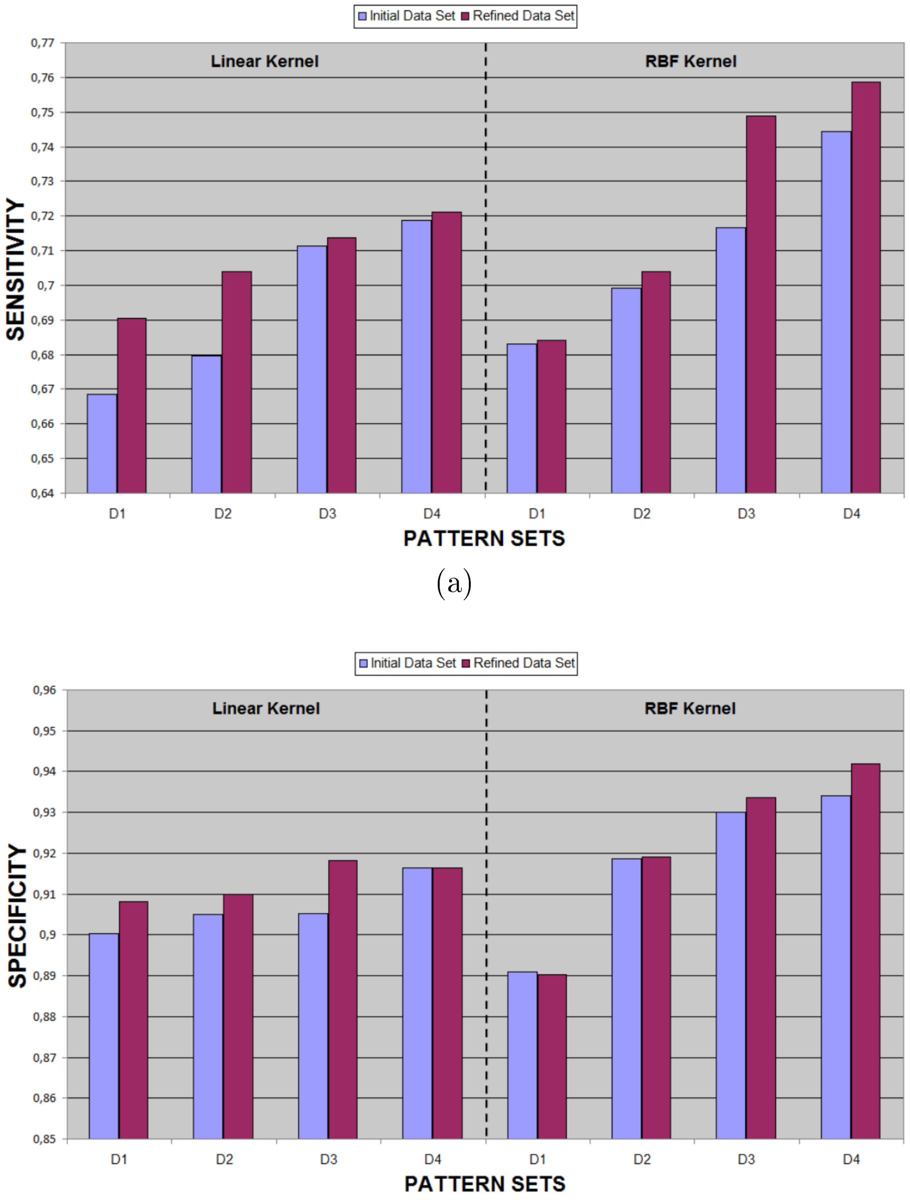
Figure 3.7: (a) The initial image with the centroids of regional minima depicted with an "×". (b) The result of the application of the distance depended rule, (c) the result of the application of the FCM, where the positive class (nuclei centroids) is depicted with a cross and the negative class (other findings) is depicted with a circle. (d) Resulted centroids of the positive class.



(a)

(b)

Figure 3.8: Results of the application of the FCM clustering with respect to (a) *Sensitivity* and (b) *Specificity* (see text for the description of the pattern sets).



(a)

(b)

Figure 3.9: Results of the application of the SVM classification with respect to (a) *Sensitivity* and (b) *Specificity* (see text for the description of the pattern sets).

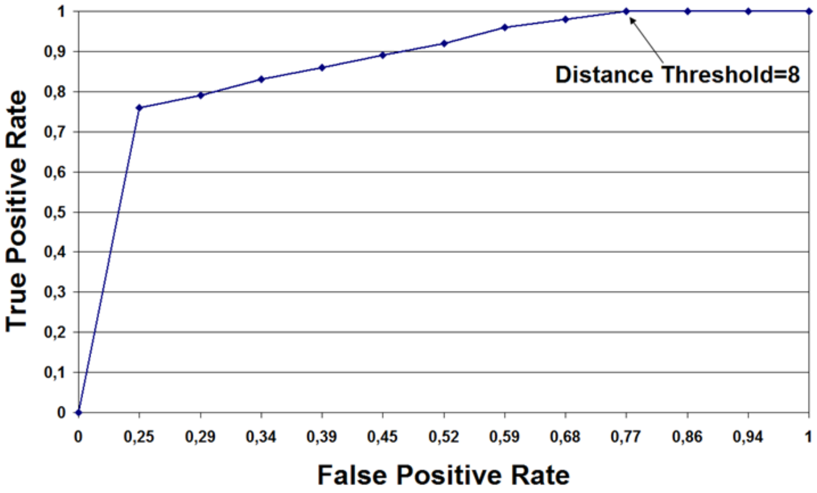


Figure 3.10: ROC curve used for the selection of the distance threshold in the distance dependent rule. Notice that for threshold values higher than 8, true nuclei centroids are missed.

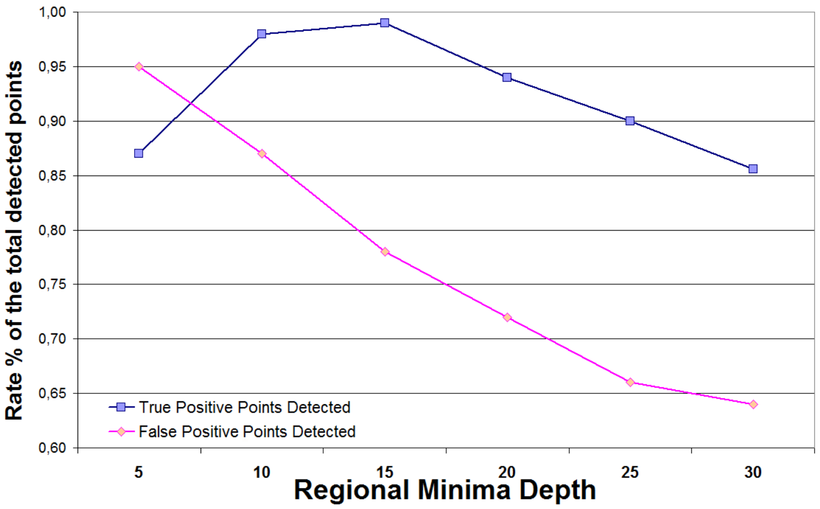


Figure 3.11: Rate of the true nuclei detected (true positives) and the false nuclei centroids detected (false positive) for different thresholds in regional minima depth. Notice that for the threshold value of 15 we obtain the maximum number of true nuclei centroids.

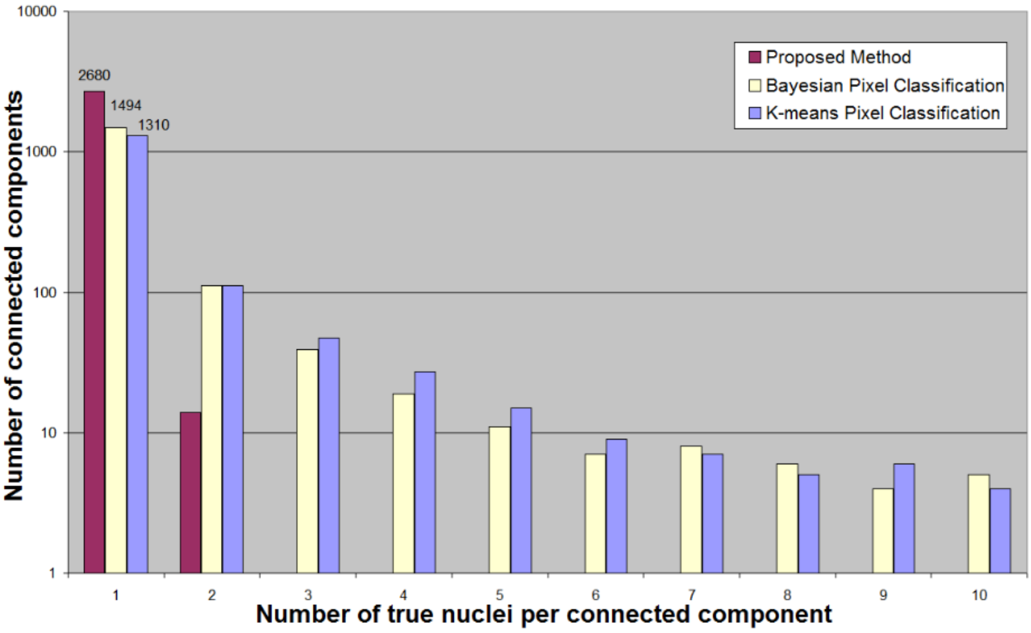


Figure 3.12: Comparative results of our method and the pixel classification schemes proposed in [1] in terms of correct nuclei localization. The index of performance is the number of connected components and the number of true nuclei (true positives) they contain. An algorithm performs well if each connected component it produces contains only one nucleus. The vertical axis has a logarithmic scale.

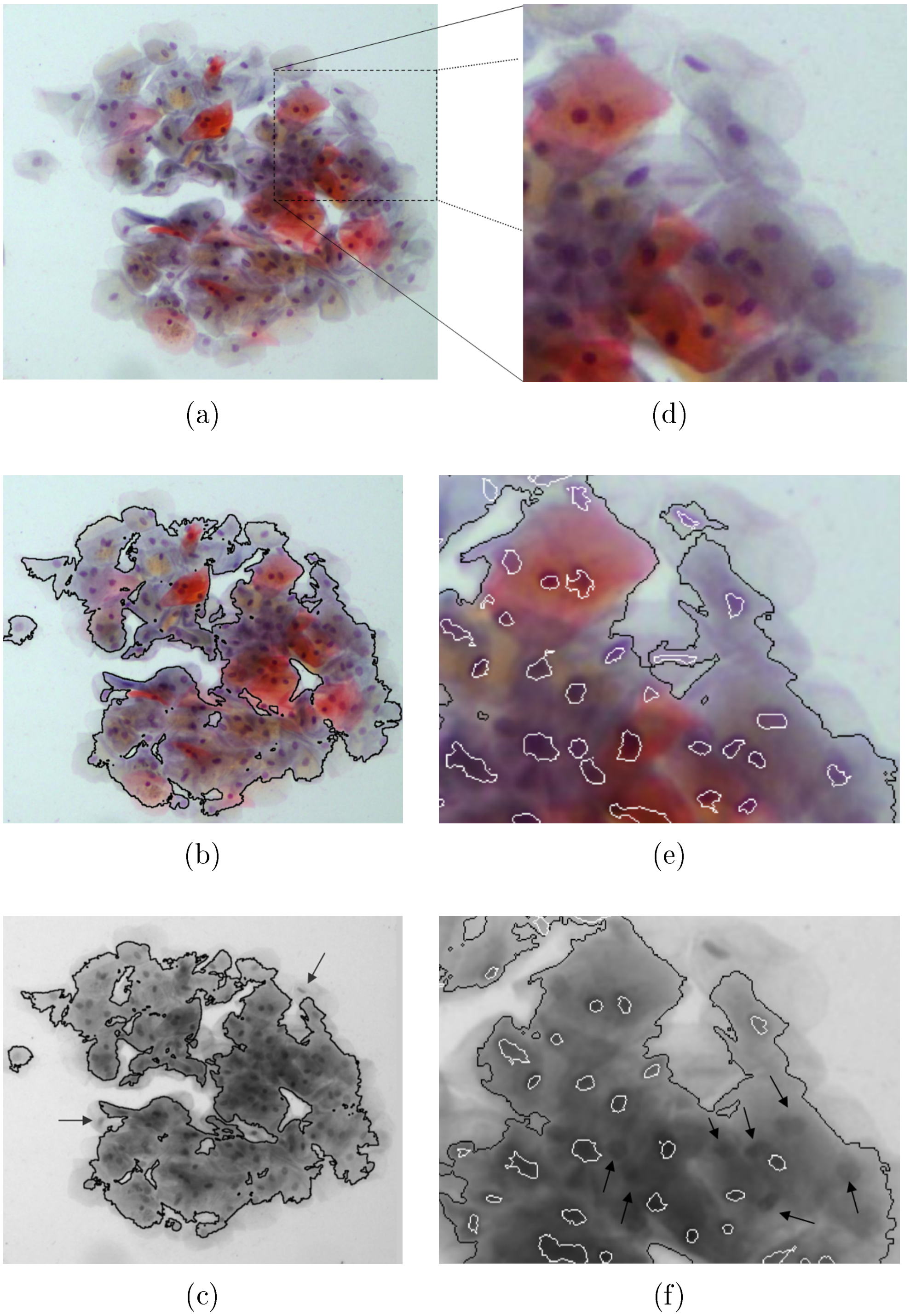


Figure 3.13: a) The initial image, (b) the result of the preprocessing step (denoted with the black line) in the color image, (c) the result of the preprocessing step in the grayscale image. The missed nuclei are marked with the arrows, (d) a part of the initial image, (e) the result of the detection of regional minima step (denoted with white lines) in the color image, (f) the result of the same step in the grayscale image. The missed nuclei of this step are marked with the arrows.

Chapter 4

Cell Nuclei Extraction by Combining Shape, Texture and Intensity Features

4.1 Introduction

The correct characterization of Pap smear slides and the derivation of conclusions for the contents of the Pap smear in a high degree depend on the general appearance of the cells nuclei. Some of the methods proposed in the literature deal only with the segmentation of the cell nucleus and cytoplasm boundaries (ignoring the detection of the nuclei position in the image). The images that are used as test set, are presegmented from the original Pap smear images and they contain only one cell and consequently one single nucleus. Several image processing methods are proposed in this scope, such as active contours [43], template fitting [40] and edge detectors [25, 26, 28]. These methods exhibit remarkable performance in the segmentation of the structural parts of the cell. However the direct application of these methods in original Pap smear images, which may contain a large number of cells, cell overlapping and image artifacts is not appropriate, as they are focused on the recognition of the boundaries of the nucleus and the cytoplasm in images which contain only one single cell.

More sophisticated approaches to the automated analysis of Pap smear images are the methods which are applied on images containing a large number of cells in cell clusters. In these methods the background is excluded and the locations and the boundaries of the cells are automatically recognized. Several approaches have been proposed, such as deformable templates [41], genetic algorithms [35], region growing with moving K-means [61] and pixel classification schemes [36]. Although these methods present promising results, their evaluation is restricted in a small data set of images and the performance criterion that is used is visual inspection, from which no reliable results about the general behavior of these methods can be obtained.

Methods based on watersheds for the analysis of Pap stained images have also been proposed in the literature. In [31], images containing one single nucleus of a Pap stained squamus epithelial cell are oversegmented with the watershed transform in order to define the differently stained subareas of the nucleus. Furthermore, in [29] watersheds are used for the detection of isolated cells in low resolution images. However, in both methods, the problem of the detection of the accurate nuclei boundaries has not been resolved. Furthermore, Lezoray *et al.* [1] proposed a method for the determination of nuclei boundaries in Pap stained serous cytologies using color watersheds, which requires the cooperation of pixel classification schemes for the extraction of the nuclei markers.

In this chapter, we describe a two-stage fully automated method for the accurate determination of the nuclei boundaries in Pap smear images, which may contain both isolated cells and cell clusters. More specifically, in the first step, nuclei and cytoplasm markers are detected using the technique described in the previous chapter. The centroids of the areas of the regional minima are considered as markers in the watershed transform for the extraction of the nuclei boundaries. The morphological color gradient image is used for the flooding process, in order to retain the color information of the image.

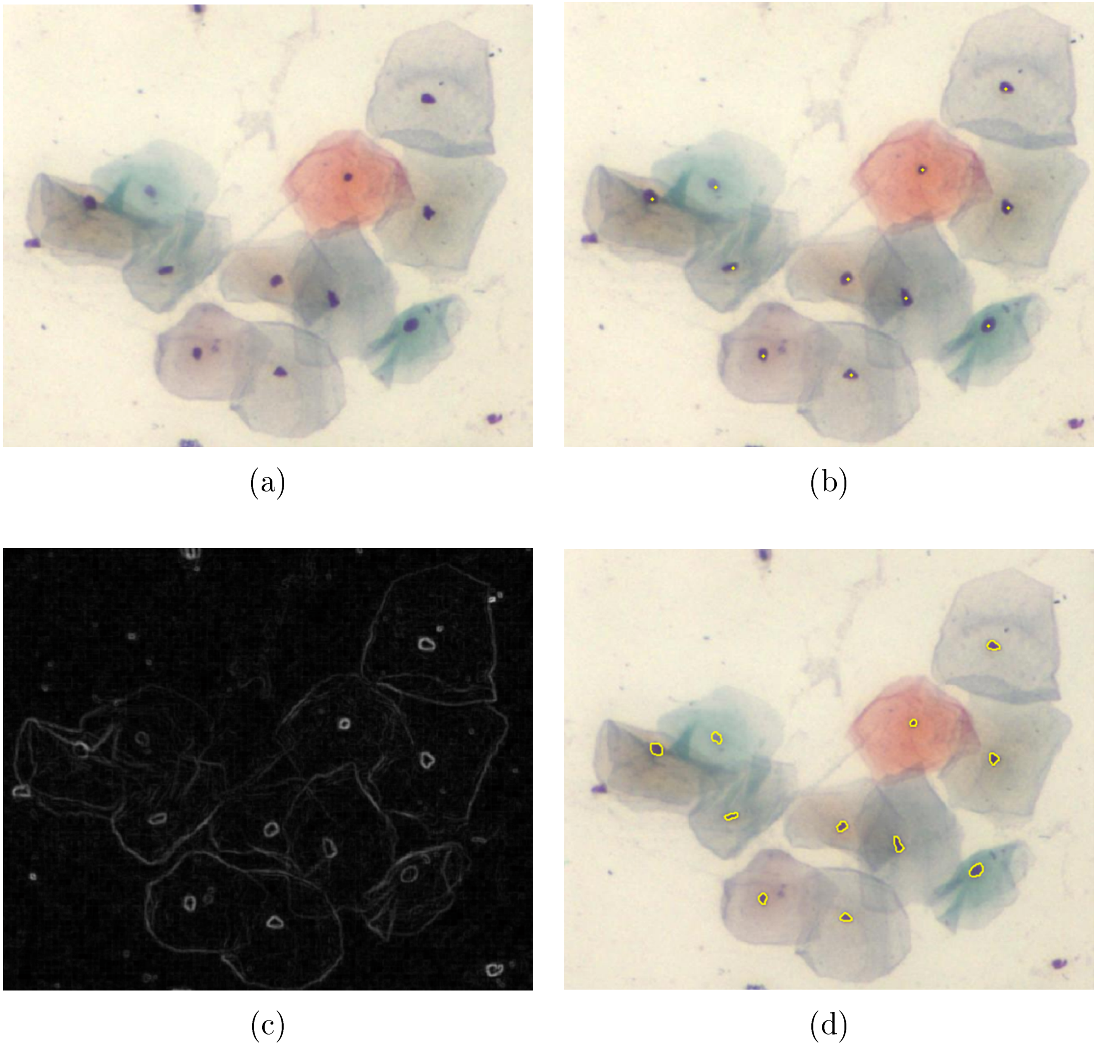
In the second stage, we extend the segmentation of nuclei boundaries with the determination of meaningful features of the detected areas, which contribute to the identification of the true nuclei in Pap smear images. It must be noted that several methods [52, 2] propose a number of cell features for the characterization of a cell as normal or abnormal. However, they involve images containing one single cell. Since our images contain overlapped cells and cell clusters, our aim is to identify the nuclei areas and to separate the results of the segmentation in two categories: the true nuclei and other findings. Therefore, from the extracted boundaries, features describing the shape and the texture of each segmented regions are calculated. In addition we have also integrate texture features and intensity disparity features of the neighborhood of each detected area. The latter evince to be some of the most discriminative features by a feature selection step based on minimum-Redundancy - Maximum-Relevance (mRMR) criterion [69]. It must be noted that in our experiments we have estimated the mRMR feature rank with two different approaches, namely using the entire image data set and the "leave-one out" strategy, as it is explained in more details in the following paragraphs.

A classification step is then performed for the reduction of unwanted findings. In this framework, the performance of two unsupervised (K-means and the spectral clustering) and one supervised (Support Vector Machine, SVM) classification schemes were examined. Our method was evaluated not only for the correct identification of cells nuclei locations but also for the accurate determination of nuclei boundaries with the boundaries obtained using the Gradient Vector Flow (GVF) deformable model [70] and a region based active contour model (ACM) [71] in terms of the Hausdorff distance from the ground truth. The method was evaluated using a large data set of 90 Pap smear images containing 10248 recognized cell nuclei, and the results indicate that the proposed method demonstrates high performance in both detection and segmentation of nuclei boundaries.

4.2 Methodology

4.2.1 Detection of the nuclei and cytoplasm markers

The first step of the proposed method is the detection of the nuclei markers in each image. This is accomplished following a two stage procedure, which includes the image preprocessing and the estimation of candidate nuclei centroids. It must be noted that the nuclei markers are obtained automatically in both isolated cells and cell clusters in the image, as it was described in section 3.2. Furthermore, we perform the distance transform in the binary mask obtained in the preprocessing step (section 3.2.1), in order to construct the cytoplasm markers.


(a)

(b)

(c)

(d)

Figure 4.1: (a) Initial image of overlapped cells, (b) the detected nuclei markers, (c) the corresponding color morphological gradient image, (d) the watershed segmentation.

4.2.2 Morphological color gradient image

For the application of the watersheds, an image containing pronounced nuclei boundaries is required. Given the fact that most of the nuclei usually have ellipse-like shape, with the intensity of the pixels inside the nucleus area lower than those lying outside, high gradient of the image across the nuclei boundaries is expected. However, the extensive variances in nuclei intensity which are present due to the staining procedure result in gradient values of nucleus/cytoplasm borders that fluctuate in a wide range. For this reason, the use of a threshold after the application of edge detectors in order to determine the nuclei edges in the image would produce noisy results, because low thresholds would result in the detection of too many false edges, while high values would result in the loss of some true nuclei boundaries (Fig. 4.2). Therefore, we construct a gradient image using the color morphological gradient [72], in order to exploit the color information of the image for the estimation of the nuclei borders. In general, the morphological gradient of a grayscale image is defined as:

$$\nabla f = \delta_g(f) - \varepsilon_g(f), \tag{4.1}$$

where $\delta_g(f)$ and $\varepsilon_g(f)$ is the grayscale dilation and grayscale erosion for a structuring element g respectively. Alternatively, the morphological gradient can be expressed as:

$$\begin{aligned} \nabla f(x) &= \max_{x \in g} \{f(x)\} - \min_{x \in g} \{f(x)\} \\ &= \max \{|f(x) - f(y)|\}, \ \forall \ \{x, y\} \in g, \end{aligned} \tag{4.2}$$

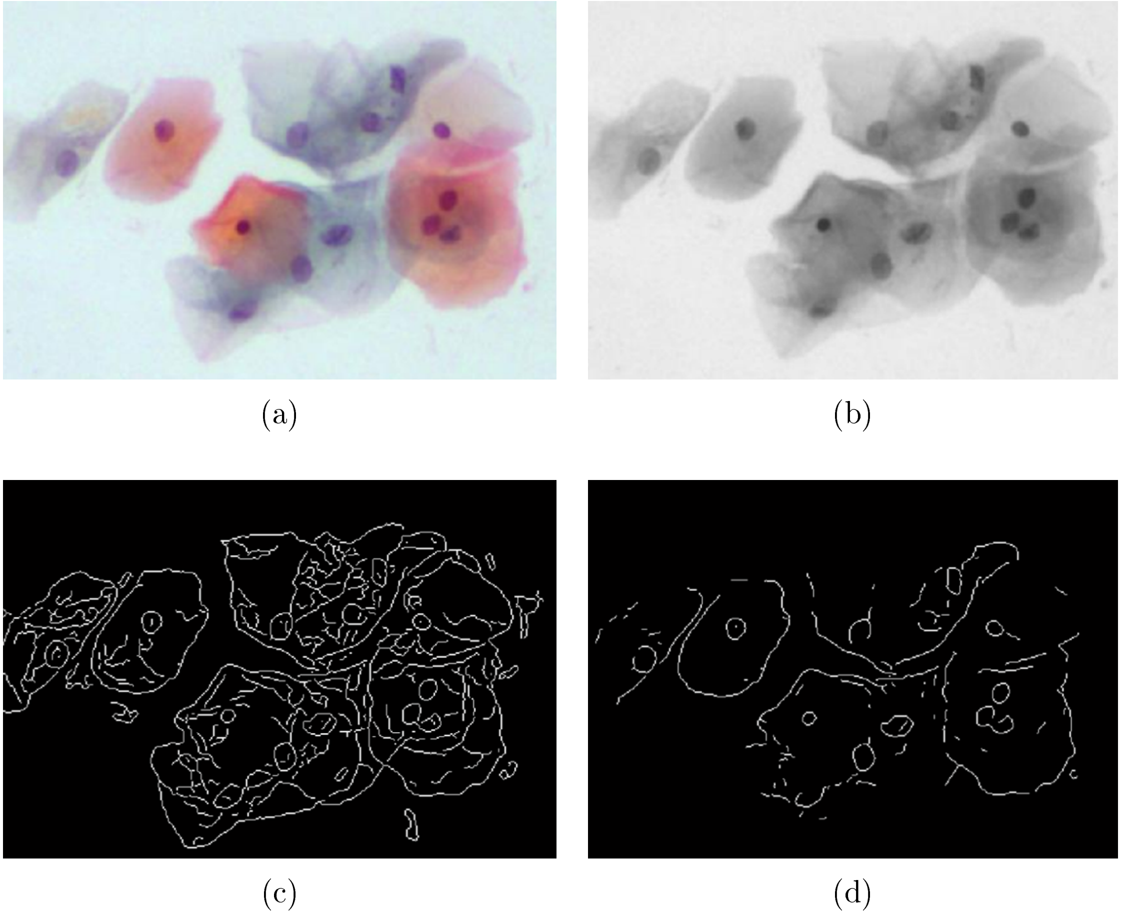
which is the maximum absolute intensity difference between two pixels in the area of the structuring element. For color images with pixels denoted as three dimensional vectors the color morphological gradient (CMG) can be expressed as:

$$CMG_g = \max_{i,j \in g} \{\|x_i - x_j\|_p\}, \tag{4.3}$$

where x_i, x_j are pixels in the structuring element g. In our experiments we compute the second norm ($p = 2$) and the structuring element that is used is a 3×3 flat structuring element. The color morphological gradient of a representative Pap smear image is depicted in Fig. 4.1(c). In this image, with the appropriate nuclei and cytoplasm markers superimposed, the marker based watershed transform (Appendix B) is applied. The result of the watershed transform in an image with nuclei markers is depicted in Fig. 4.1(d).

4.2.3 Clustering of the candidate nuclei

The determination of the watershed lines, usually results in the correct identification of the nuclei positions in the image. However, some false positive areas are also detected, due to the existence of a regional minimum. This is a consequence of the detection of the nuclei markers step, which produces some centroids of regional minima that do not indicate the existence of nuclei (Fig. 4.3). Therefore, the elimination of these areas is necessary and a clustering step is performed for the separation of the detected areas into



(a)

(b)

(c)

(d)

Figure 4.2: (a) Initial image of overlapped cells and (b) the corresponding grayscale image, in which we apply the Canny edge detector. Using a small threshold results in (c) an image with many undesired edges, while using a high threshold results in (d) an image with several significant edges missing.

two classes: the true nuclei class and the rest of the findings. Thus, for every detected area a vector of features is determined, which will be used as input to the clustering algorithms.

Feature extraction

The efficient separation of the true nuclei regions from the total segmented regions requires the generation of meaningful features of very good discriminative ability. Having found the areas of the nuclei enclosed by the detected boundaries, features concerning the shape, the texture and the intensity of the detected regions can be easily determined. However, the restriction of the calculation of these features only for the area enclosed by the detected boundaries is not sufficient because regions of regional minima not corresponding to true nuclei may also have similar features. In this step it is expedient to take advantage of the fact that the nuclei are darker than the surrounding cytoplasm and the detected nuclei regions would present significant differences from their neighborhood. Moreover,

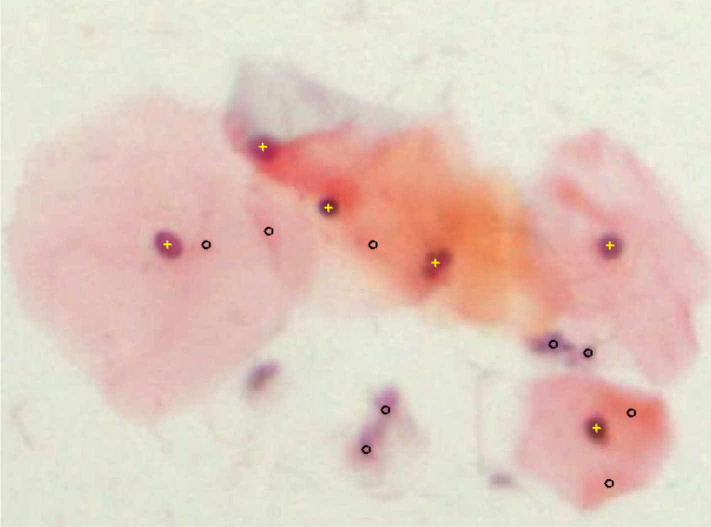


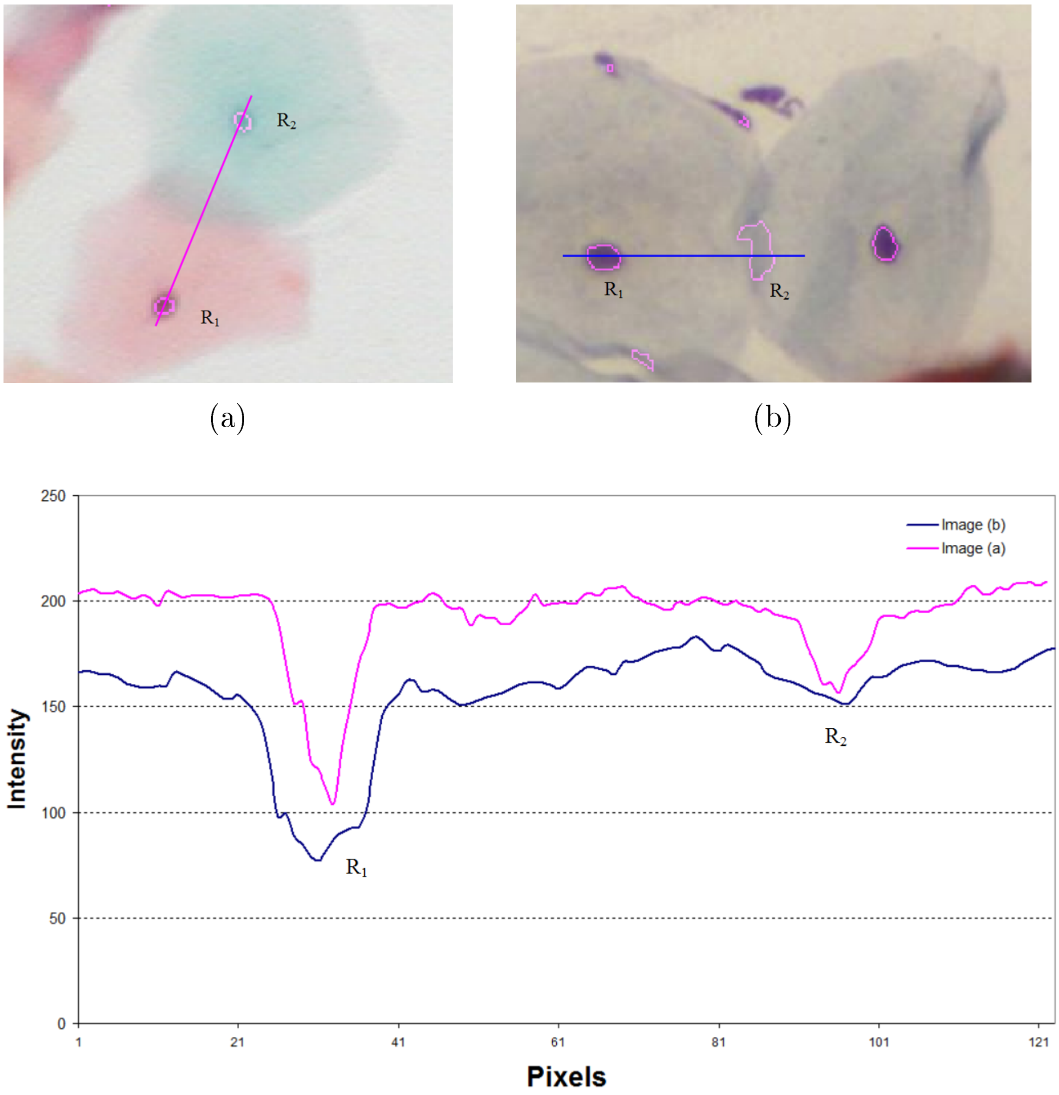
Figure 4.3: The detected centroids of the regional minima in the image. The true nuclei locations are represented by a yellow cross and the false positive findings are represented by a black circle.

the detected regions that do not belong to nuclei were probably detected due to the existence of shallow minima in the intensity of the area of the cytoplasm in the image, and they are more likely to present similar features values from their neighborhood (Fig. 4.4).

For this reason, we propose the calculation of features also for the neighborhood of the detected areas, which is defined in terms of the bounding box of these areas (Fig. 4.5). More specifically, for each detected area A, the bounding box B is calculated as the maximum rectangle that contains the detected region, and the neighborhood Ngh is determined as the complement A^c in B, that is $Ngh = A^c \cap B$. In our work, for the construction of the feature set, the pixels within the detected region, the pixels of the neighborhood and the pixels of the bounding box are taken into account. Three categories of features are developed: shape, textural and intensity disparity features.

Shape Features The detected boundaries for the nuclei are expected to present an ellipse-like shape and several features to describe this characteristic are chosen. More specifically, six features extracted from the shape of the detected region boundary are calculated, that is the Circularity, the Eccentricity, the Major and the Minor Axis Length, the Equivalent Diameter of a circle with the same area as the region, and the Perimeter of the detected region. The Major Axis Length, the Minor Axis Length and the Eccentricity are defined in terms of an ellipse that has the same central second moments as the region. The shape features are presented in (Table 4.1).

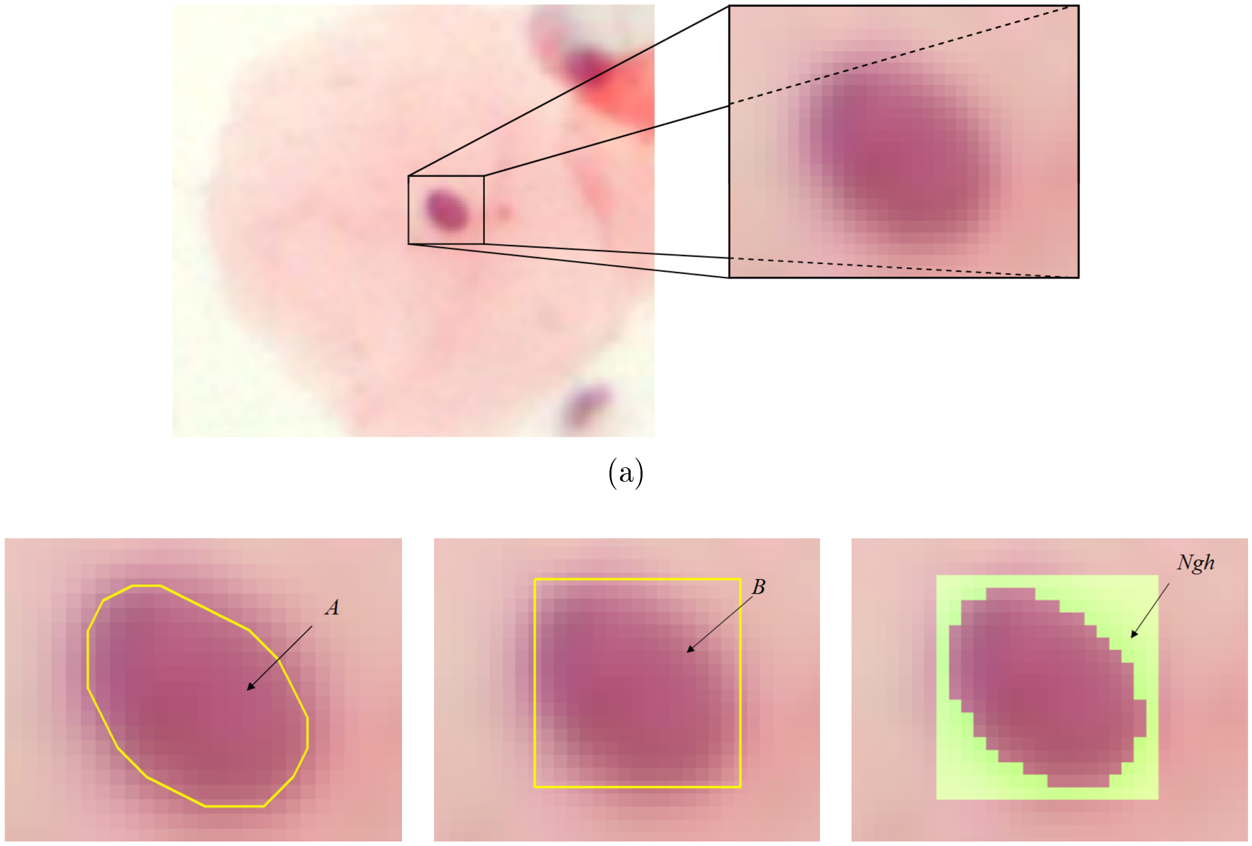
Textural Features The texture analysis of the detected regions is based on the statistical properties of the intensity histogram in the three color components and the



(a) (b)

(c)

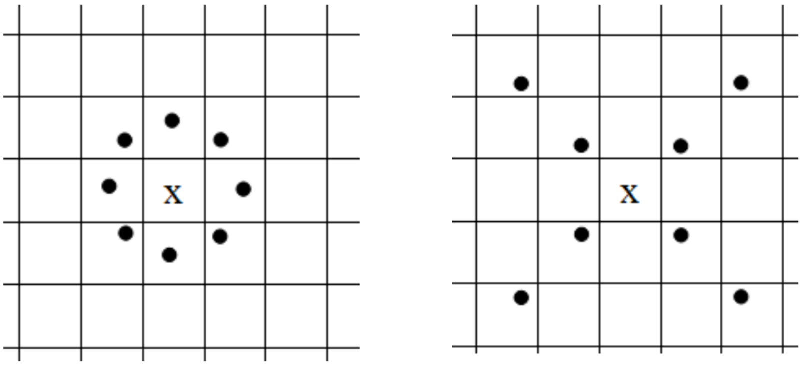
Figure 4.4: (a)-(b) The result of the watershed transform in parts of two different cell images. The regions R_1 and R_2 that are detected in both images with the watershed transform are joined with a line for better visualization purposes. In (a), the detected areas R_1 and R_2 correspond to the areas of true nuclei, while in (b), the detected area R_1 corresponds to a nucleus and the area R_2 corresponds to a cytoplasm overlapping area. The variation of the average color image intensity value along the line which joins the areas R_1 and R_2 is depicted in (c). Notice that for the area R_1 we observe sharp reduction of the intensity value in both images. For the area R_2, although the average intensity value is similar in both images, sharper intensity reduction (in relation with its neighborhood pixels) occurs only for the true nucleus in image (a). This indicates that the use of the neighborhood of each detected area contributes in the recognition of the true nuclei.



(a)

(b) (c) (d)

Figure 4.5: The selected areas for the construction of the feature set. (a) A cell from the initial image, (b) the detected nucleus boundary with the watershed transform and the enclosed area A, (c) the area B of the bounding box of the detected boundary, (d) the area of the neighborhood Ngh $(A^c \cap B)$ of the detected nucleus.



(a) (b)

Figure 4.6: The topology of the neighborhood used for the calculation of the LBP [2, 3]: (a) circle, (b) hyperbola.

Table 4.1: Shape Features.

Minor Axis Length*	$K = \sqrt{\frac{2(u_{20}+u_{02}-\Delta)}{u_{11}}}$
Major Axis Length*	$L = \sqrt{\frac{2(u_{20}+u_{02}+\Delta)}{u_{11}}}$
Eccentricity	$E = 2\frac{\sqrt{(\frac{L}{2})^2-(\frac{K}{2})^2}}{L}$
Equivalent Diameter	$ED = \frac{4\times Area}{\pi}$
Perimeter	P =number of boundary points
Circularity	$C = \frac{4\pi\times Area}{P^2}$

*The formulas for Δ and the central moments u_{pq} of order $p+q$ of the region $s(x,y)$ are defined as: $\Delta = \sqrt{4u_{11}^2+(u_{20}-u_{02})^2}$ and $u_{pq} = \sum_x \sum_y (x-\bar{x})^p(y-\bar{y})^q$, where $\bar{x}$ and $\bar{y}$ are the coordinates of the centroid of the region.

calculation of some texture descriptors such as the local binary patterns (LBP, see Appendix C) [2, 3]. Thus, for every segmented region we have calculated the Third Moment, the Uniformity, the Entropy and the Smoothness of the intensity histogram for the three predefined regions (A, B, Ngh). Moreover, the normalized uniform rotation-invariant LBP occurrence histogram was calculated for the bounding box (B) of the segmented regions, using LBP of two different neighborhood topologies: a circle (LBP_c^{riu2}) of unit radius and a hyperbola (LBP_h^{riu2}) with semi-major and semi-minor axis lengths equal to one (Fig. 4.6). In both topologies, the number of equally spaced pixels was $P = 8$ (see Appendix C for more details and [2] and [3] for a more in depth explanation of these features). The mean and the standard deviation of each histogram were used as features. All the textural features are calculated for all three color channels and they are summarized in (Table 4.2).

Intensity Disparity Features The feature that characterizes the intensity of each region is the average of the intensity value of all the pixels of the region. However, as it is observed, the average intensity of the nuclei varies in a wide range and may coincide with regions of cell overlapping in the image. An equivalent intensity feature that pronounces the disparity of the detected region and its neighborhood is the difference of the average intensity between those regions (Table 4.3). We expect high values for this feature when it refers to nuclei regions, as the nuclei area is darker than the surrounding cytoplasm. Three values of this feature were calculated independently for the red, green and blue component of the original image.

Table 4.2: Texture Features.

Third Moment*	$\mu_3 = \sum_{i=1}^{L-1} (z_i - m)^3 p(z_i)$
Uniformity	$U = \sum_{i=1}^{L-1} p^2(z_i)$
Entropy	$e = -\sum_{i=1}^{L-1} p^2(z_i) log_2 p(z_i)$
Smoothness	$R = 1 - \frac{1}{1+s^2}$, $s = \sqrt{\sum_{i=1}^{L-1} (z_i - m)^2 p(z_i)}$
Mean LBP_c^{riu2}	See Appendix C
StDev LBP_c^{riu2}	See Appendix C
Mean LBP_h^{riu2}	See Appendix C
StDev LBP_h^{riu2}	See Appendix C

*Given that z_i is the intensity value i and $p(z)$ is the histogram of the intensity levels in a region with L possible intensity levels, then the average intensity of the region is calculated as $m = \sum_{i=0}^{L-1} z_i p(z_i)$

Table 4.3: Intensity Disparity Features.

Foreground-Background contrast in red*	$dR = m_{RED}^{Ngh} - m_{RED}^{A}$
Foreground-Background contrast in green	$dG = m_{GREEN}^{Ngh} - m_{GREEN}^{A}$
Foreground-Background contrast in blue	$dR = m_{BLUE}^{Ngh} - m_{BLUE}^{A}$

* m_{color}^{region} is the average intensity value of an image region in a specific color component. The RGB color space is used in our experiments and the regions of the image that are considered are the enclosed boundary area A and its neighborhood $Ngh = A^c \cap B$, where B is the bounding box of area A.

Feature selection

For each detected region we have calculated in total 57 features. More specifically, 6 features concerning the shape of the region, 3 features concerning the intensity disparity of the detected areas and their neighborhood and finally, for the three color components, 3×4 textural features for the enclosed area (A), 3×4 textural features for the neighborhood

$(Ngh), 3 \times 8$ textural features for the bounding box B were calculated. However, the contribution of each feature is different in the categorization of the data. For the selection of the most discriminative features, a feature selection technique is employed which is based on the Minimum-Redundancy-Maxium-Relevance (mRMR) criterion [69]. More specifically, given a data set of N samples of M features $X = \{x_i^j,\ i = 1,..,M,\ j = 1,..,N\}$, and the target classification variable c, the objective is to find from the M dimensional space R^M a subset of m features that characterizes c more efficiently.

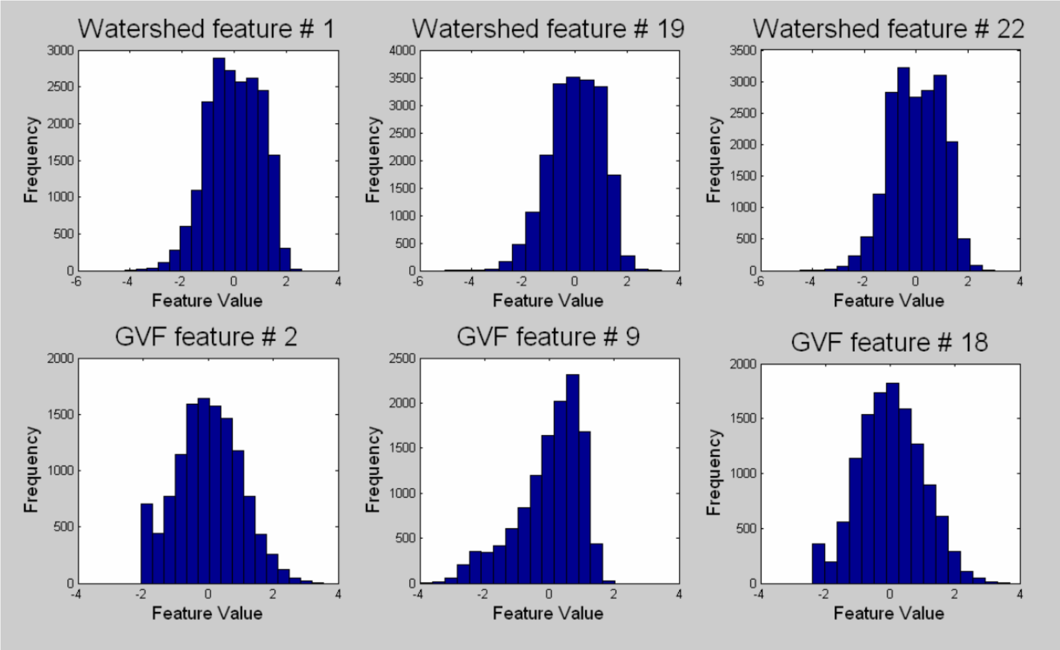


Figure 4.7: Representative histograms of some features of the watershed and the GVF segmentation. Notice that their distribution consists of a single blob and this allows their discretization into three states at the positions $\mu \pm \sigma$.

The mRMR criterion combines both Max-Relevance (max$D$) and Min-Redundancy criteria (min$R$), which are defined respectively as [69]:

$$\max_{S \subset X} D(S, c), \quad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c), \tag{4.4}$$

$$\min_{S \subset X} R(S), \quad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j), \tag{4.5}$$

where S is the feature set and $I(x; y)$ is the mutual information between two random variables, which is defined in terms of their marginal and joint probability density functions $p(x), p(y)$ and $p(x, y)$ as:

$$I(x; y) = \int \int p(x, y) log \frac{p(x, y)}{p(x)p(y)} dxdy. \tag{4.6}$$

The mRMR criterion is then defined as:

$$\max_{S \subset X} \left(D(S,c) - R(S) \right). \tag{4.7}$$

The selection of features for the construction of the final set is obtained incrementally, that is if $m-1$ features are already selected in the S_{m-1}, then the m^{th} selected feature will be the one that satisfies eq. (4.7). The optimal size of the features set depends on the specific classification algorithm that will be used.

Thus the features were ranked in a range beginning from the most powerful discriminative feature to the feature with the least discriminative power. It must be noted that for the calculation of the mutual information, each feature variable was discretized into three states at the positions $\mu \pm \sigma$ (μ is the mean value and σ is the standard deviation of the specific feature distribution). More specifically, it takes -1 if the feature value is less than $\mu - \sigma$, 1 if the feature value is larger than $\mu + \sigma$ and 0 otherwise. This assumption is reliable when our features follow a unimodal-like distribution. This was verified by the construction of the histograms of each feature and some representative examples are depicted in Fig. 4.7. In Table 4.4 the first 16 most discriminative features for all the segmentation techniques are presented.

Application of classification/clustering algorithms

For comparison purposes, three clustering methods are employed for the separation of the detected areas in the true nuclei class and the other findings class: the K-means [73], the spectral clustering [74] (see Appendix A) and the Support Vector Machine (SVM) classifier with the radial basis function (RBF) kernel [63]. Given the fact that the K-means and the spectral clustering algorithms do not require any training, they are applied independently in each image. However, for the application of the SVM classification algorithm a training data set is constructed. In our experiments, we use the "leave one out" technique for the evaluation of the performance of the classifier. Thus, 21 slides were used as training set and the remaining slide was used as test set. This experiment was repeated 22 times, each time using a different slide as test set. The performance of the classification is calculated using the trained SVM classifier in the test set.

4.3 Results

4.3.1 Study group

We have collected 90 images from 22 different Pap stained cervical cell slides, which were acquired through a CCD camera adapted to an optical microscope. We have used a 10× magnification lens and the acquired images of size 1536×2048 were stored in JPEG format. The total number of cell nuclei in the images, which were identified by an expert observer is 10248. In order to obtain the ground truth, the nuclei locations were manually identified.

Table 4.4: mRMR rank of the 16 most discriminative features for the watershed, the GVF and the ACM segmentation in decreasing order with respect to their discriminative ability. The features highlighted in bold face are common for all segmentation methods.

	Watersheds	GVF	ASM
1.	Entropy of B in green	**Foreground-Background contrast in green**	**Foreground-Background contrast in red**
2.	Perimeter	Minor Axis Length	Minor Axis Length
3.	**Foreground-Background contrast in red**	Third moment of A in blue	Uniformity of Ngh in green
4.	**StDev LBP_h^{riu2} in green**	**StDev LBP_h^{riu2} in red**	**StDev LBP_h^{riu2} in red**
5.	**Circularity**	Entropy of Ngh in red	Smoothness of Ngh in green
6.	**Foreground-Background contrast in green**	Mean LBP_c^{riu2} in green	Eccentricity
7.	Mean LBP_c^{riu2} in blue	**Foreground-Background contrast in blue**	**Foreground-Background contrast in green**
8.	Entropy of B in red	Eccentricity	Mean LBP_c^{riu2} in blue
9.	**Mean LBP_h^{riu2} in blue**	**Mean LBP_h^{riu2} in blue**	**Mean LBP_h^{riu2} in blue**
10.	Smoothness of B in green	Uniformity of B in green	Third moment of A in red
11.	StDev LBP_c^{riu2} in red	**Foreground-Background contrast in red**	**Circularity**
12.	Entropy of A in green	StDev LBP_c^{riu2} in red	**Foreground-Background contrast in blue**
13.	**Foreground-Background contrast in blue**	**StDev LBP_h^{riu2} in green**	**StDev LBP_h^{riu2} in green**
14.	**StDev LBP_h^{riu2} in red**	**Circularity**	Entropy of Ngh in red
15.	Smoothness of A in red	Entropy of B in green	Mean LBP_c^{riu2} in red
16.	Third moment of Ngh in blue	Third moment of Ngh in red	Third moment of A in blue

4.3.2 Numerical evaluation

The presented method was bilaterally evaluated in order to estimate the performances of the clustering algorithms for the detection of the true nuclei in the images, and also the accuracy of the segmentation, in comparison with the ground truth (manually traced nuclei boundaries). Furthermore the method performance was compared with the corresponding performance of two different segmentation techniques, namely the GVF deformable model [70] and the ACM model [71] in terms of both classification and segmentation results. In the detected regions of both GVF and ACM segmentation techniques, the previously described features were determined. For the classification performance, we have calculated the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) findings in all images of our data set, and we have defined the *Sensitivity* and the *Specificity* statistical measures. In addition, the segmentation performance was evaluated with the calculation of the Hausdorff Distance ($D_{Hausdorff}$) between the manual traced boundary M and the boundary Ψ obtained from the segmentation procedure defined as:

$$D_{Hausdorff} = \max_{\alpha \in M} \left\{ \min_{b \in \Psi} \{D(\alpha, b)\} \right\}, \tag{4.8}$$

where D is the Euclidean distance.

It must be noted that in the detection of the nuclei markers the method misses in total 147 true nuclei position which is a total loss rate of 1.01%. Thus in the following steps, the total number of true nuclei is reduced to 10101.

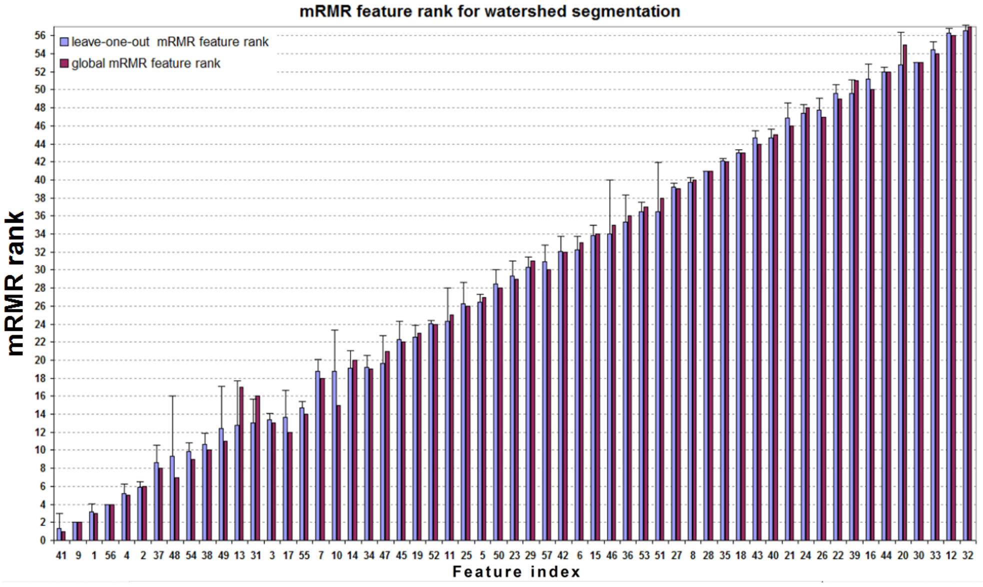


Figure 4.8: The leave-one-out and global mRMR feature rank for the watershed transform. For the leave-one-out mRMR feature rank the standard deviation is also depicted with error bars.

Classification

In our experiments we have tested several configurations of the classification process which involve both the calculation of the mRMR feature rank and the clustering algo-

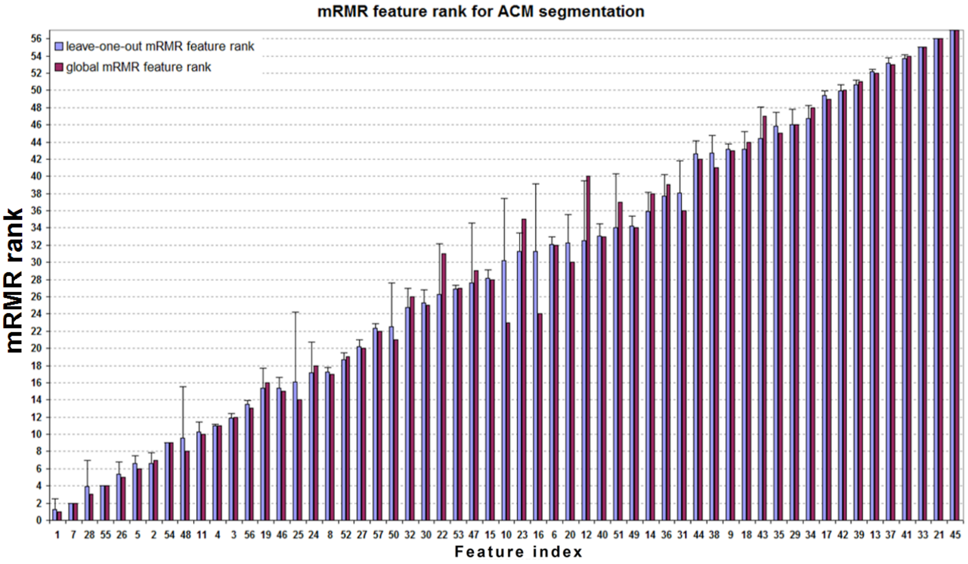


Figure 4.9: The leave-one-out and global mRMR feature rank for the ACM segmentation algorithm. For the leave-one-out mRMR feature rank the standard deviation is also depicted with error bars.

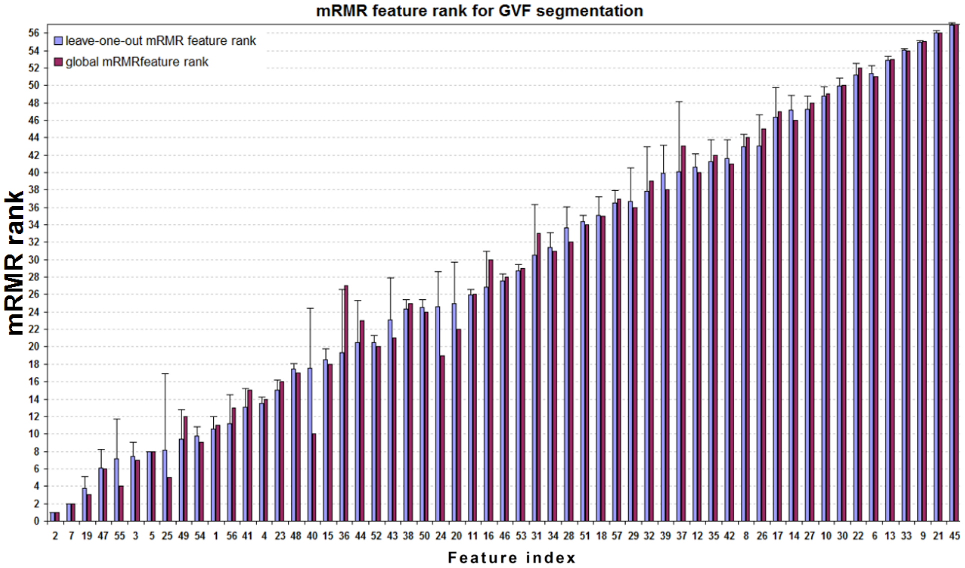


Figure 4.10: The leave-one-out and global mRMR feature rank for the GVF segmentation algorithm. For the leave-one-out mRMR feature rank the standard deviation is also depicted with error bars.

rithms (K-means, spectral clustering and SVM). For this reason we include the following experiments, which were executed for all the data sets obtained from the three segmentation algorithms (watersheds, GVF, level sets). The estimation of mRMR rank of feature was determined in two different ways:

1. The estimation of mRMR rank of feature was determined in two different ways:

   - The whole data set of patterns was used as input to the mRMR criterion (global mRMR) and a ranking was obtained which was then used in the classification algorithms.
   - The set of patterns was separated into 22 folds (each fold corresponds to a single slide). Then, 21 folds were used for training and the remaining fold was used for testing. From the training set, we obtained the mRMR rank (leave-one-out mRMR) of features and this rank was used in the classification algorithms applied to the image of the testing slide. This procedure was repeated 22 times, each time using a different slide as test set. By these means, we obtained 22 different feature ranks, which were assigned in the 22 folds (slides).

   Therefore, all of the classification techniques (K-means, spectral clustering and SVM) were executed twice, using the above mRMR rankings (global mRMR and leave-one-out mRMR). For the selection of the ideal number of features, the performance of the classification techniques was estimated on the test set using a pattern of increasing dimension varying from 2 to 57 features. Starting from a pattern described by only 2 features, one feature was added incrementally until all of the 57 features are employed. In the second case described above, the selection of the feature that is added in the pattern is different for each test slide (and consequently for the images belonging to this slide) and it was determined by the corresponding mRMR rank (obtained using the other 21 slides as training set). In order to evaluate the importance of each feature, the mean position and its standard deviation in a feature histogram was calculated (Figures 4.8, 4.9, 4.10).

2. The estimation of the best value for parameter σ in spectral clustering was also obtained using a leave-one-out strategy. The set of patterns was separated into 22 folds (each fold corresponds to a single slide), with 21 folds were used for training and the remaining fold was used for testing. Several experiments with different values for σ were performed in the training set, using patterns containing all of the features (the dimension of each pattern was 57). Then, we selected the value of σ that exhibited the best performance in the training set. This value was used for the application of spectral clustering in the images of the test set. This procedure was repeated 22 times, each time using a different training and test set.

3. The values of the parameters of the SVM classifier (γ and C for the RBF kernel) were obtained by constructing two different data sets, each one containing half of the slides (11 slides were randomly selected for the training set and the remaining were used as test set). We performed several experiments with different pairs of values for γ and C, namely $(C, \gamma) \in \{0.01, 0.125, 0.25, 0.5, 1, 2, 4, 8\}$, while the SVM classifier was trained with the training set of patterns containing 57 features. Afterwards the performance of the classifier was estimated with the test set. The values for γ and

C were selected as those which exhibit the best performance of the SVM classifier in the test set and they were $\gamma = 0.01$ for all the segmentation methods and $C = 2$ for the GVF segmentation and $C = 4$ for the watershed and the ACM segmentation.

The number of features that results in the best classification performance depends on the specific classification algorithm. When a performance criterion is maximized for a specific number of features, then this subset of features is selected. In our work, the performance criterion that the clustering algorithm should maximize is the harmonic mean (HM) of the *Sensitivity* and the *Specificity* defined as:

$$HM = \frac{2 \times Sensitivity \times Specificity}{Sensitivity + Specificity} \tag{4.9}$$

Recall that these quantities were defined in (3.13) and (3.14).

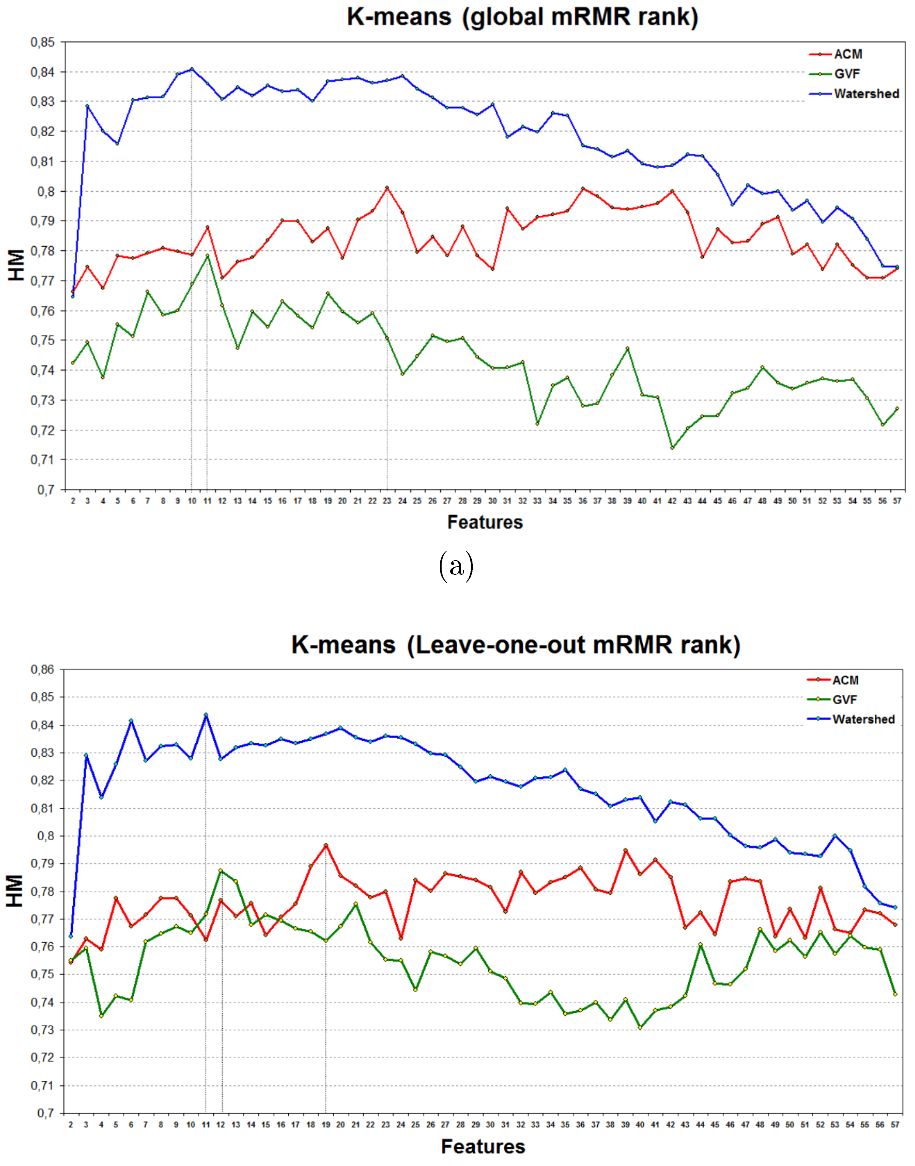
In Fig. 4.11, the values of HM criterion versus the number of features are depicted for the ACM, GVF and the watershed segmentation for the K-means algorithm. Similar experiments were performed for the definition of the best feature subset using the spectral clustering algorithm (Fig. 4.12) and the SVM classifier (Fig. 4.13). The performance of the SVM classifier for the watershed and the GVF segmentation increases as more features are used, and reaches the maximum performance at 57 features. For the ACM segmentation, the SVM classifier reaches the maximum performance at 26 features. In all cases as it can be observed, the HM measure for the watershed segmentation is higher than the other two segmentation techniques. More specifically, the best results in terms of the HM for all the segmentation schemes using the global and the leave-one-out mRMR rank are presented in 4.5. As we can see, the best results were obtained with the K-means clustering algorithm using patterns obtained from the watershed segmentation. The SVM classifier is preferable for the ACM and GVF segmentations, as it produced higher performances than the K-means and the spectral clustering. Furthermore, in most of the cases, the use of leave-one-out mRMR feature rank produces better results in comparison with the use of the global mRMR rank. It must be noted that for comparison purposes, the performance of the SVM classifier was selected for 26 features for all the segmentation techniques.

Table 4.5: Clustering Performance.

	K-means		Spectral clustering		SVM	
mRMR type	Global	Leave-one-out	Global	Leave-one-out	Global	Leave-one-out
Watersheds	84.09%	84.36%	82.64%	82.93%	82.46%	82.52%
ACM	80.09%	79.64%	76.84%	77.00%	81.87%	81.95%
GVF	77.83%	78.76%	77.20%	77.33%	80.20%	80.28%

Segmentation

In order to evaluate the performance of the segmentation method, the obtained nuclei boundaries were compared with the corresponding resulted nuclei boundaries of the GVF

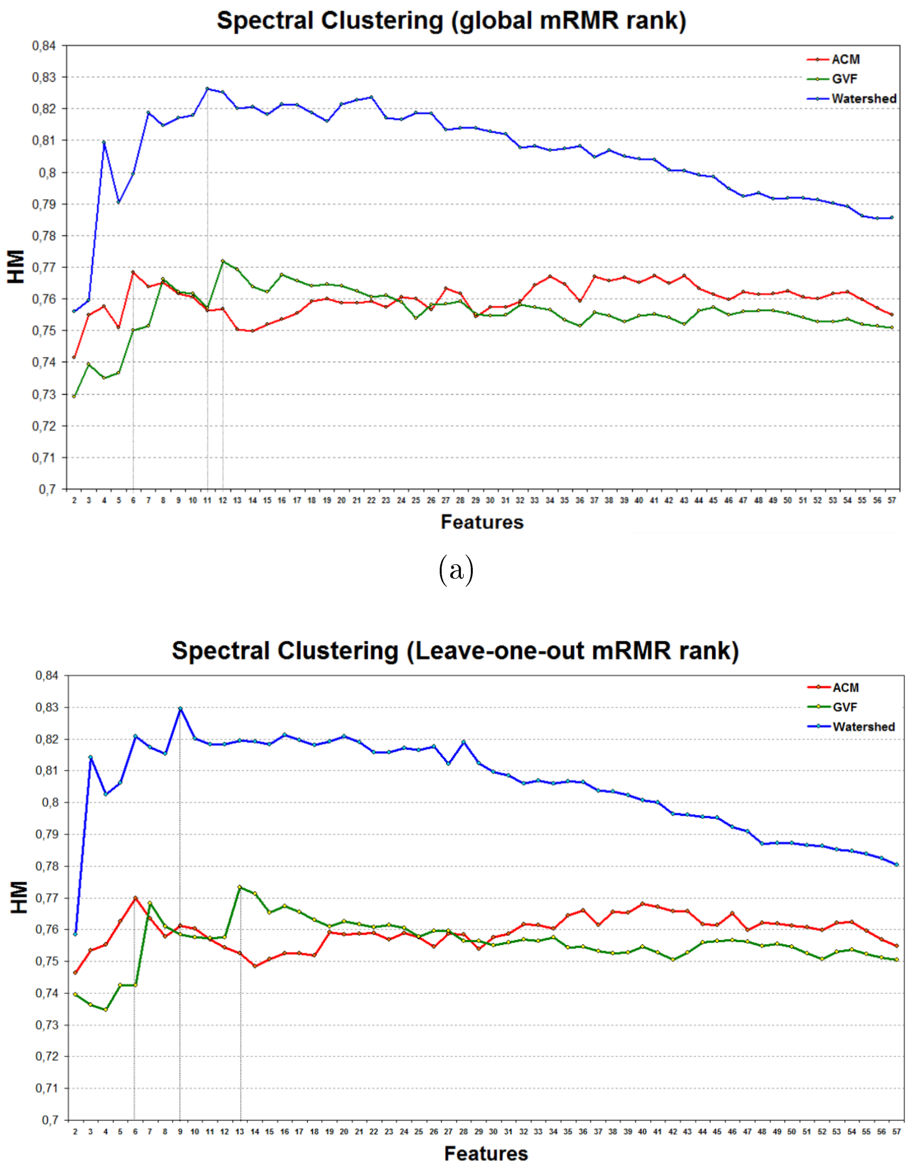


(a)

(b)

Figure 4.11: Results in terms of the HM measure for the K-means clustering for ACM, GVF and watershed segmentation for both (a) global and (b) leave-one-out mRMR rank. The vertical line indicates the number of features where the HM measure takes its maximum value for the three segmentation methods. These values of HM are contained in Table 4.5.

deformable model and the ACM model and also with the manually traced boundaries. It must be noted that for the application of the GVF deformable models, an initial approximation of every nucleus boundary is required. For this reason, we search for some points in the neighborhood of each detected centroid, which are likely lying in the nucleus circumference [75]. In the morphological color gradient image, having as starting points the candidate nuclei centroids we construct a circular searching grid with

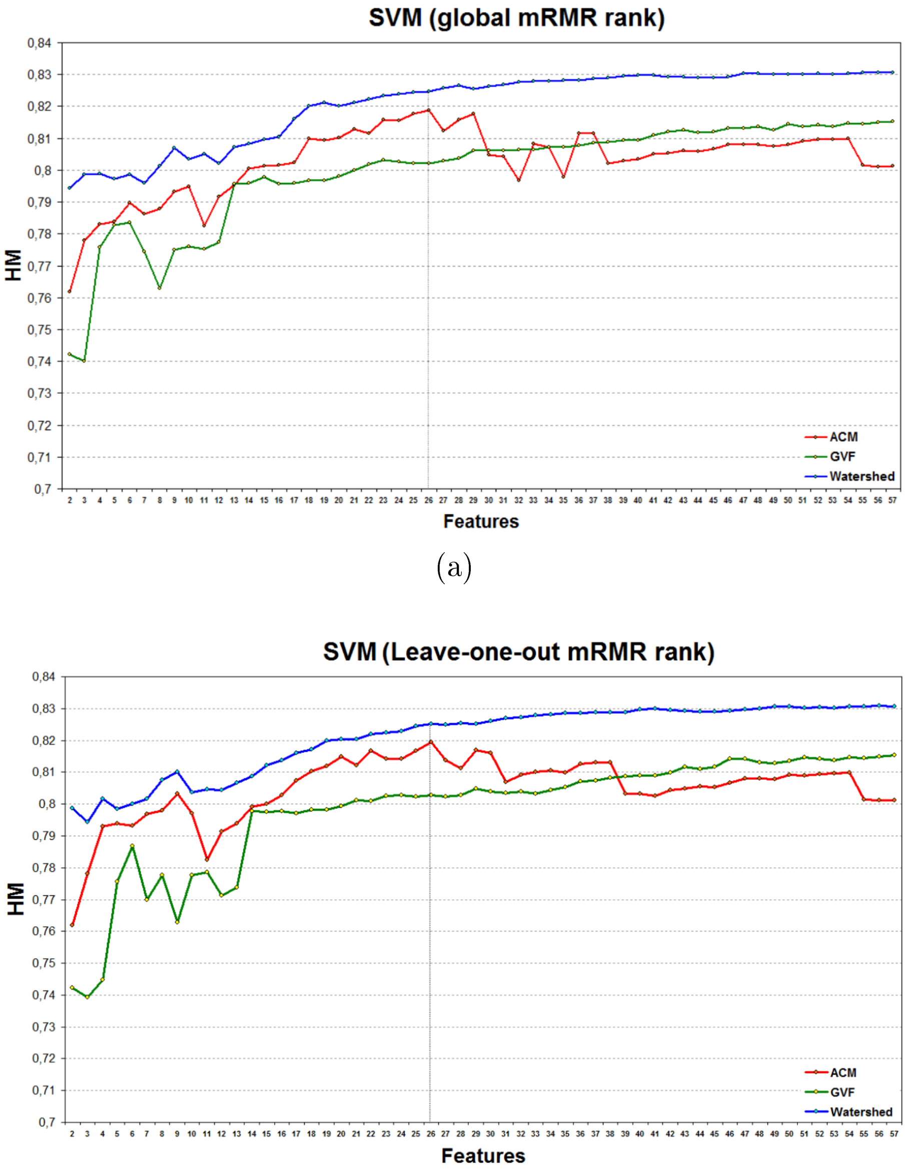


(a)

(b)

Figure 4.12: Results in terms of the HM measure for spectral clustering for ACM, GVF and watershed segmentation for both (a) global and (b) leave-one-out mRMR rank. The vertical line indicates the number of features where the HM measure takes its maximum value for the three segmentation methods. These values of HM are contained in Table 4.5.

8 radial profiles consisted of 8 points each and centered at the location of each candidate nucleus centroid. In each radial profile we choose the pixel with the highest intensity (non maximum suppression) and the initial approximation of the nuclei boundaries is obtained with the convex hull of the circumferential points found in the this step. The values for the weighting parameters of the GVF deformable model are fixed for all the images and they are set to be $\alpha = 0.9$ for the tension, $\beta = 1.5$ for the rigidity and $\gamma = 3$ for the image



(a)

(b)

Figure 4.13: Results in terms of the HM measure for the SVM classification for ACM, GVF and watershed segmentation for both (a) global and (b) leave-one-out mRMR rank. For comparison purposes, the indicative values for HM measure were evaluated using the first 26 features. The values of HM are contained in Table 4.5.

force.

In a similar way, the ACM model was also applied to the same images. More specifically, having found the nuclei markers, we apply the ACM model, as it is described in [71] in the 21×21 image window centered at each marker. The model was initialized as a rectangle in the middle of the selected neighborhood and it was applied in the morphological color gradient image with $\alpha = 20$, where α is the balloon force which controls the contour shrinking or expanding.

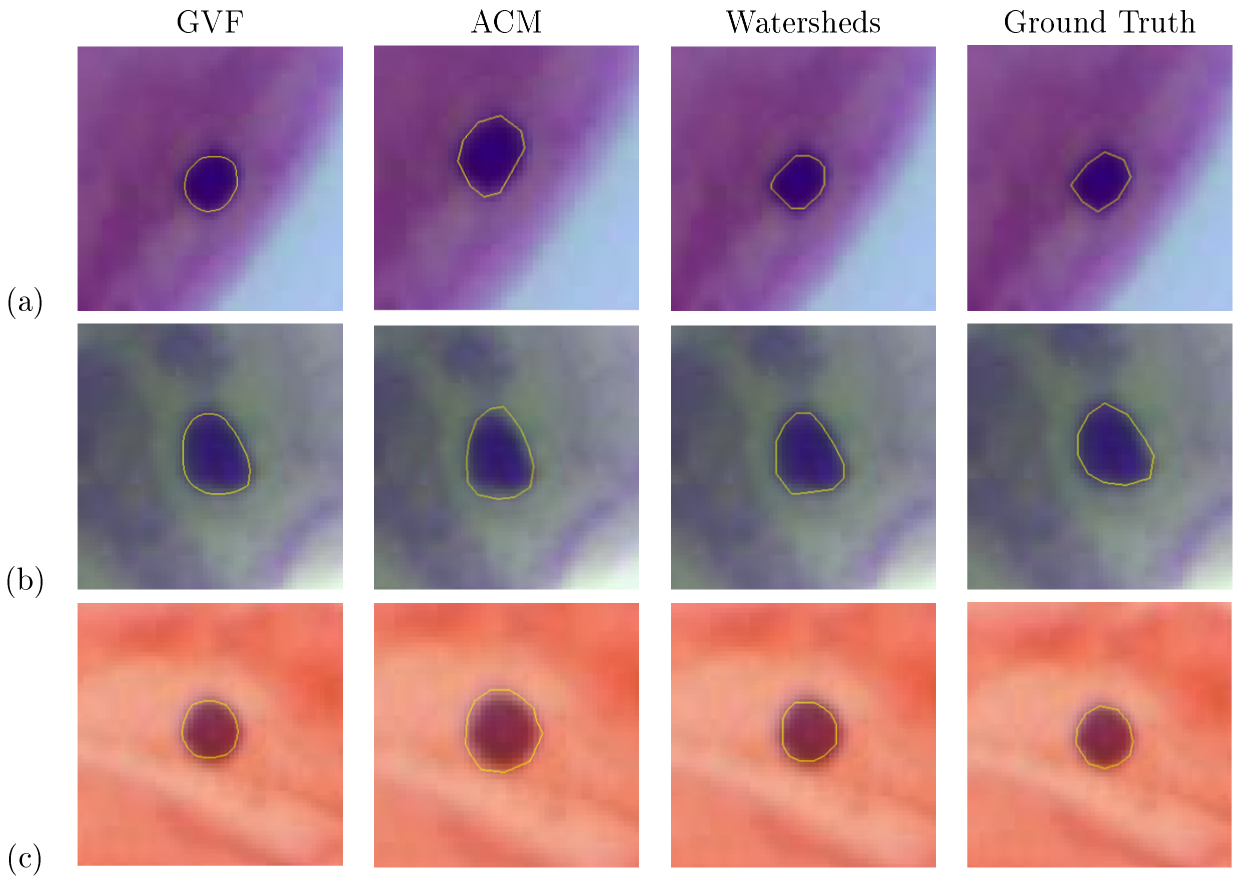


Figure 4.14: (a)-(c) Segmentation results for several detected nuclei.

Several examples of the segmentation results are depicted in Fig. 4.14. The Hausdorff Distance for the ground truth and the watershed segmentation was estimated as 1.71 ± 0.54 ($\mu \ \pm \ \sigma$). The corresponding distance for the GVF and ACM segmentation is 2.65 ± 3.23 and 2.48 ± 2.30 respectively. This implies that the watershed segmentation is closer to the manually traced nuclei boundaries, and as a result it is more accurate than GVF and the ACM segmentation. Furthermore, the ACM segmentation is more performing than the GVF segmentation, as it exhibits lower Hausdorff Distance. In the next paragraph, some reasons of failure for the GVF and ACM segmentations are discussed.

4.4 Discussion

The proposed method for the segmentation of the cell nuclei in Pap smear images is fully automated and it can be applied directly in any conventional Pap stained cervical smear images, in order to produce accurate nuclei boundaries. It consists of five steps: the preprocessing, the estimation of the candidate nuclei centroids, the application of the watershed transform, the feature extraction and the classification step. The method was developed in Matlab using a dual core PC with a 2.0 GHz processor and 3GB of RAM. The execution time for each step of the method depends on several factors, such as the proportion of the image characterized as background in the preprocessing step, the number of the candidate nuclei centroids in each image, the classification algorithm and the number of features in each pattern. An indicative execution time for the segmentation of the images (steps one to four of our method) is 2-5 min. The mean execution time of K-means in an image using 16 features is less than a second, while the corresponding time for the spectral clustering algorithm is 5-6 seconds. Finally, the mean execution time for the training of the SVM classifier using 21 slides and the evaluation of the performance in the test set (one slide) varies from 2 to 4.5 minutes.

The parameters used in the several steps of the segmentation method were determined after careful examination of the images by an expert cytopathologist in combination with the results of several tests. For the detection of the nuclei markers, the same parameters are used, as they are described in the previous chapter. However, as it was mentioned before, this step misses some of the true nuclei positions. This is mainly due to the faintly staining and the uneven layering of some cells. In the first case, the cells are undistinguished from the background and as a consequence, the nuclei of these cells are considered as isolated objects in the image background and they are rejected as image artifacts. In the second case, the intensity of the nucleus does not well differentiate from the cytoplasm intensity and no regional minimum is detected in the nucleus position.

The nuclei markers obtained in the previous step are used in the application of the watershed transform. The importance of this step is crucial, as it prevents from the oversegmentation that would be produced by the application of the watershed transform in the images without markers. Hence, using the detected cytoplasm markers, the flooding process starts from a position in the catchment basins of the nuclei area and finally converges to the actual boundaries of the true nuclei. Furthermore, the problem of the detection of false positive detected centroids is effectively resolved in the classification step.

The feature selection using the mRMR criterion produces different feature ranks for the three segmentation techniques, as it can be observed in Table 4.4. This is a consequence of the differences between the segmented regions provided by each method (ACM, GVF, watersheds) necessitating different features for its representation. As it can be observed by the feature ranking, we can conclude that the discriminative ability of some features is equally important for all the segmentation techniques, as seven of them were selected by all of the segmentation techniques in the first 14 positions. These features are highlighted in bold face fonts in Table 4.4. Furthermore, from figures 4.8, 4.9, 4.10 we can observe that in general, the standard deviation of the features selected by the leave-one-out mRMR is rather insignificant for the first 10 and the last 20 positions in the mRMR rank, which indicates that from the entire data set of features, the most discriminative and the least discriminative features are the same for every fold (slide) of our image data set.

As it was verified by the results, the watershed segmentation is more accurate than

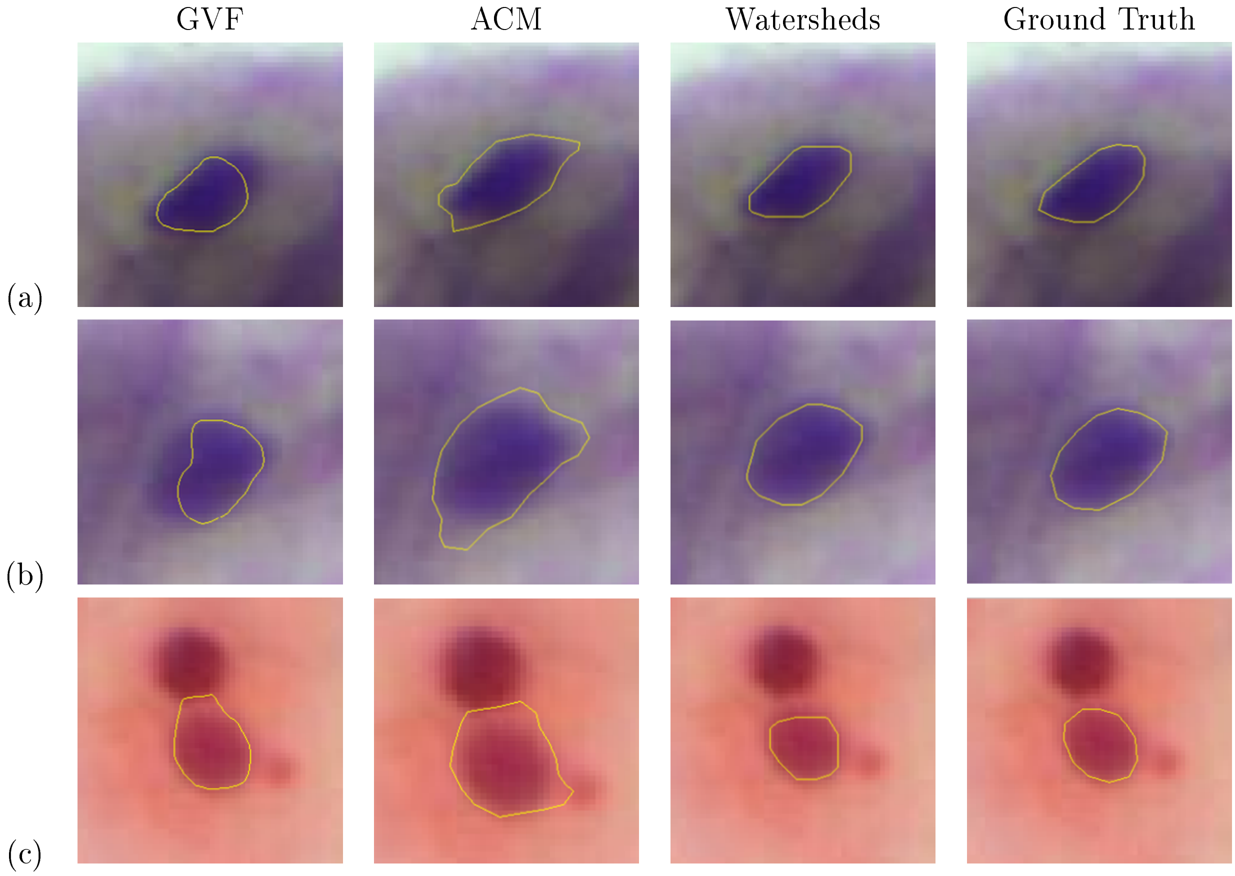


Figure 4.15: Representative cases of failure for ACM and GVF segmentation in images with (a) weak gradient at the nucleus boundary, (b) the inhomogeneities of the nucleus intensity and (c) the existence of high value of gradient in the neighborhood of the nucleus boundary.

the GVF and the ACM segmentation. For both these segmentation techniques, the main reason of failure is that their behavior highly depends on the values of their parameters. Furthermore, the existence of a high gradient value in a small distance of the detected nucleus and the inhomogeneities on the nuclei intensity affect the performance of these techniques. Some examples of these cases are depicted in Fig. 4.15. In Fig. 4.15(a) as the gradient in the border of the nucleus/cytoplasm is weak, the shape of the GVF deformable model is mainly determined by its internal forces, which enforce it to be of a relatively small length and smooth. Furthermore, in Fig. 4.15(b) the existence of intensity variations in the area of the nucleus attracts the points of the GVF deformable model, which converges to a position far from the actual nucleus boundary. For the same images, the ACM model also fails to accurately determine the nucleus borders. Finally, in Fig. 4.15(c) both the GVF and ACM model are attracted by the points of high image gradient, which do not correspond to the boundary of the detected nucleus. In all these cases the GVF and ACM model do not succeed in detecting the accurate nucleus boundary. In contrast, as it is observed, the watersheds overcome these limitations and produce nuclei boundaries that are closer to the ground truth.

The accurate determination of the nuclei boundaries leads to the calculation of more accurate features, which improve the performance of the clustering algorithms. This is the reason why the use of features extracted with the watershed segmentation present better classification performance than the corresponding features extracted from the GVF and the ACM segmentation. Furthermore, for the determination of a feature set we exploit the fact that the true nuclei area presents significant variations with respect to its neighborhood and the calculation of neighborhood features would result in the effective discrimination of the true nuclei areas and the false positive areas. This is also confirmed by the use of mRMR criterion (Table 4.4), which indicates that for the feature set obtained with all the segmentation techniques (Watersheds, GVF, ACM), at least 7 out of 10 most discriminative features concern the outer area (bounding box B and neighborhood Ngh) of the detected boundaries.

Traditionally, immediate fixation and staining of the cellular sample on the slide with 70% ethyl alcohol and Papanicolaou stain have been established as the professional standard. This fixation and staining combination results in a cellular sample, that, not only has well-defined and tinted morphological features, but also its transparency allows for microscopic visualization of nuclear and cytoplasmic boundaries through multiple layers of epithelial cells. In our work, we use 90 conventionally stained Pap smear images, which exhibit several differences in colorization (e.g. the blue color can vary from deep blue to light blue). Although we have not included any process of color correction and detection of improper staining, the method provides accurate results when it is applied to the images of our data set.

4.5 Conclusion

The identification of the cervical cell nuclei areas is a prerequisite for the derivation of diagnostic conclusions and the characterization of the contents in the Pap smear images. The automated detection and segmentation of the nuclei boundaries in these images is a challenging issue, as these images present several limitations. In this work, we have effectively overcome the problem of the detection of the nuclei locations and we have developed a fully automated method for the segmentation of cell nuclei in Pap smear images. Moreover, we propose the determination of a meaningful feature set for the detected areas, which results in the efficient discrimination of the true nuclei class by the clustering algorithms. As it is verified by the results, the method produces more accurate nuclei boundaries which are closer to the ground truth, compared to the GVF deformable model and the ACM segmentation method. The main advantage of the proposed method is that it can be applied directly in Pap smear images obtained by an optical microscope, without any observer interference, for the accurate automated identification of the cell nuclei boundaries.

Chapter 5

Overlapping Cell Nuclei Segmentation Using A Spatially Adaptive Active Physical Model

5.1 Introduction

One of the most interesting and challenging issues in the automated analysis of microscopic images, is the delineation of the overlapped cells or cells nuclei. The cell overlapping areas in the microscopic slides are very common phenomena, especially in the case of the well-known Pap smear. The segmentation of these images has been studied by several researchers [26, 43, 76, 77], as the nucleus is the structural part of the cell that presents significant changes when the cell is affected by a disease. Furthermore, features that are based on the nuclei shape have been used by several researchers for the discrimination of normal and abnormal nuclei [52, 55]. Thus, the accurate and detailed identification of the nucleus shape is important for the correct interpretation of the Pap smear.

Usually, the existence of the different layers of the cervical specimen in the slide results in areas, where the cells of an upper layer partially obscure the cells lying underneath. In real time microscopic examination this problem is commonly solved in most cases with the adjustment of the lens focus, and the cells of different layers are clearly identified. However, in static images acquired through a digital camera adopted on a microscope, this is not possible and for this reason efforts have been made by several researchers in order to contribute to the automated segmentation of the overlapped cells or overlapped nuclei in many cytological images.

More specifically, in terms of the general segmentation techniques used for the separation of overlapped nuclei, the geometric active contours are used in [78]. In this approach, each cell is represented by its own level-set function and a coupling constrain prevents neighboring contours from overlapping each other and maintaining the separation of similar cells in contact. Furthermore, the distance transform in a binary image containing the regions of the nuclei is calculated in [5] and the topographic surface generated by the distance transform is considered as a Gaussian mixture. The EM algorithm is then applied for the determination of the parameters of each nucleus cluster and the separation line of the overlapped nuclei is defined with the minimization of a criterion function. It must be noted, that in this approach, the occluded area of each nucleus is reconstructed with a constrained ellipse fitting technique.

The segmentation technique that has been extensively used in many methods for the separation of clustered cell nuclei is the watershed transform [79, 4, 80, 81]. The main concern for these methods is to overcome the oversegmentation resulting by the watershed transform. For this purpose, special attention has been paid on the determination of marking strategies and the selection of appropriate nuclei and background markers. Thus, shape markers are extracted using an adaptive H-minima transform and a marking function based on the outer distance transform is introduced in a watershed-like algorithm in [79] for the separation of clustered nuclei. In [4], the segmentation of clustered nuclei is treated as an optimization problem and a marker extraction scheme based on the H-minima transform is introduced to obtain the optimal segmentation result from the distance map. In order to optimize the selection of markers from which the flooding process will start during the watershed-based segmentation, prior information about the usual shape of cells nuclei, which is included in the determination of two templates for the identification of aggregating or overlapping nuclei, is employed in [80]. Finally, in [81], the boundary of the overlapping nuclei is extracted through the marker based watershed transform and the separation of touching cells is obtained by ellipse fitting.

The above methods were applied in different cytological images such as fluorescence in situ hybridization (FISH) images [79, 80, 82] or microscopic images from several specimens such as mammary invasive ductal carcinoma or cervical images [5, 4]. In the first case, the separation line between the clustered nuclei is obtained through the application of the watershed transform. However, these cases are very sensitive to the selection of appropriate markers, in order to identify the correct location of each nucleus marker and a marker for the overlapping area between the nuclei. Furthermore, in [4, 78, 81], the partial nuclei boundaries lying in the areas of overlap are estimated with ellipse fitting algorithms, exploiting the prior knowledge about the elliptical shape of the nuclei.

In this chapter, we present an alternative method for the separation of overlapping nuclei which is based on the representation of the nucleus shape by the vibrations of a spring-mass system [83] and the statistical learning of the vibration modes of the system in the framework of Active Shape Models (ASM) [84]. More specifically, through physics-based shape parameterization, the elastic 2-D boundary modeling can be achieved by a closed chain topology of virtual masses on the contour. Each node of the model has a mass and it is connected with two neighboring nodes through springs with the same stiffness and damping parameters. The physics-based equations of motion govern the deformation of the model [83], which reaches the equilibrium when it is placed on the object's boundary. Segmentation methods based on these models have been proposed for multimodal brain image analysis [85], skeleton family generator [86] and reconstruction of serially acquired slices for the determination of volumes [87].

In addition, active shape models (ASM) [84] are well-known parametric deformable models which are based on the construction of a statistical model of the global shape variation from a training set of shapes. They have been extensively used for the recognition and localization of objects that follow the same geometric form of a sample of well known shapes, such as face detection [88], biomedical image segmentation [89] and handwritten character recognition [90]. Image segmentation with ASM requires the representation of the shape of the object of interest by a set of points. Based on this representation, a deformable model is iteratively deformed to fit to an instance of the object of interest in an unknown image. The model is constrained by the Point Distribution Model (PDM) [84], in order to vary only in ways that are learnt in a training set of labeled examples.

Our work combines the segmentation of an image with ASM [84] and the representation of an object using modal analysis [83]. Thus, a physical model is adopted in the training phase, in which the parameters to be learnt are the variations of the modes of the model. The attributes of the nuclei shapes are expressed in terms of modal analysis and in the training phase the modal distribution is estimated. Therefore, a more compact description of the shape model is obtained. Next, we develop a framework for the deformation of an active physical model similar to ASM, for the detection of an unknown new nucleus in images containing two overlapped nuclei.

It must be noted that the idea of combining the physically-based and the active shape models was first proposed in [91]. However, no closed form solution for the deformation of the model was provided, and only admissible deformations in the neighborhood of the initial position of the model were acceptable, providing a heuristically obtained solution for the deformation of the model. Furthermore, in [41] a method for the segmentation of nuclei boundaries based on the elliptical approximation of the nucleus boundaries and a deformable model, which can accommodate a certain degree of variability is presented. However, in that work the motion of the deformable model is controlled by a probabilistic framework and the parameters that control the variability of the shapes have been selected experimentally.

As it will be explained in the following paragraphs, our method significantly differs from the aforementioned work in many perspectives. First of all, we provide a closed form solution for the deformation of the model, which does not depend on trial and error based admissible configurations in the shape of the model, but it is based on the dynamic change of the generalized displacement, in order the model to be attracted from the significant characteristics of the image and also to be consistent with the learnt parameters. Thus, the generalized displacements are updated through gradient based optimization. Furthermore, the model converges to a position close to the desired boundary, as it always provides admissible solutions. This is possible even in the case of the overlapping part of the nuclei where no significant edges are present, as the contribution of the external energy in the model deformation is mutable.

The method proposed herein is motivated by [85]. The main difference is that in [85] the *a priori* knowledge was obtained for one anatomical structure (the skull) and then, the other structures (e.g. the brain) are estimated by least squares. In our case, the *a priori* knowledge concerns the occluded parts of the nuclei, whose estimation relies on an iterative gradient descent method resulting naturally from the linear form of the model.

Moreover, there are two main differences of our method with respect to standard ASM algorithm: first, the modal amplitudes of the learnt model are used instead of the 2D landmark points and second, the cost function to be minimized for the detection of the nuclei boundaries is affected by the locally adaptive image force, which is introduced in order to extract reliable nuclei boundaries in the regions of overlap. An important characteristic of the proposed method is that it provides a smooth representation of the nucleus boundary by the physical model, which entails in the reduction of the number of parameters employed in the segmentation step. This is possible since the number of the modal amplitudes that contain the most important information about the shape of the model is small, due to the principal components analysis step. Furthermore, another significant characteristic is that the proposed model is flexible and it converges fast in the position of the desired boundary, due to the linear transformation adopted in the deformation step. Finally, the method is evaluated using a test set of 50 cytological images of conventional Pap smears, which contain two overlapped nuclei each and it presents high performance, as it is verified by the results.

5.2 Methodology

5.2.1 Training phase

In this phase, the physical model is constructed and it is applied on a training set of images containing a single nucleus, in order to learn the modal distribution describing the shape of the nuclei. The basic steps of the training phase are described in detail in the following paragraphs.

Construction of the physical model

In order to obtain a compact representation of the shape of the nuclei boundary, we adopt the physical deformable model proposed by Nastar and Ayache [83]. A physics based deformable model is used, whose behaviour is controlled by the governing equations of motion. More specifically, the physical model consists of N virtual masses located at points $\mathbf{X}(t) = \{\mathbf{x}_1(t), \mathbf{x}_2(t), , ..., \mathbf{x}_N(t)\}$. The motion of the physical model towards the border of the object of interest is expressed by a finite element formulation and is estimated by solving a $2N$-dimensional differential matrix equation (for the horizontal and vertical direction):

$$\begin{aligned} \mathbf{M}\ddot{\mathbf{U}}_x(t) + \mathbf{C}\dot{\mathbf{U}}_x(t) + \mathbf{K}\mathbf{U}_x(t) = \mathbf{F}_x(t) \\ \mathbf{M}\ddot{\mathbf{U}}_y(t) + \mathbf{C}\dot{\mathbf{U}}_y(t) + \mathbf{K}\mathbf{U}_y(t) = \mathbf{F}_y(t) \end{aligned} \tag{5.1}$$

where $\mathbf{M}$, $\mathbf{C}$ and $\mathbf{K}$ are $N \times N$ matrices describing the mass, the damping and the stiffness of the model. Moreover, $\mathbf{F}_x$ and $\mathbf{F}_y$ are vectors containing the image force at the nodes locations and $\mathbf{U}_x$, $\dot{\mathbf{U}}_x$, $\ddot{\mathbf{U}}_x$ and $\mathbf{U}_y$, $\dot{\mathbf{U}}_y$, $\ddot{\mathbf{U}}_y$ are the vectors of displacement, velocity and acceleration of the model in the horizontal and vertical direction respectively.

The above equations describe the equilibrium between internal and external forces of the system. The internal forces are expressed by the definition of the virtual masses of the model and the interaction between them, while the external forces are usually defined as the intensity or the gradient of the image at the pixels where the nodes of the model are located. The system (5.1) can be solved by setting the initial values of displacement and velocity equal to zero and then using an explicit Euler scheme. However, instead of solving directly the equilibrium equation (5.1), we can use a frequency based technique called modal analysis, which describes the motion of the model in terms of the free vibrations of the system.

More specifically, at a first step the following change of basis is used [85]:

$$\mathbf{U} = \mathbf{\Phi}\tilde{\mathbf{U}} \tag{5.2}$$

where $\mathbf{\Phi}$ is a square non-singular matrix and $\tilde{\mathbf{U}}$ is the vector of the generalized displacement. The columns of the matrix are selected to be the eigenvectors of the generalized eigenproblem:

$$\mathbf{K}\phi_i = \omega_i^2 \mathbf{M}\phi_i \tag{5.3}$$

where ϕ_i is the i-th mode and ω_i its frequency. This is an effective way for the expression of the displacement vector $\mathbf{U}$ in terms of modal displacements, that is :

$$\mathbf{U} = \mathbf{\Phi}\tilde{\mathbf{U}} = \sum_{i=1}^{N} \tilde{u}_i(t)\phi_i \tag{5.4}$$

where $\tilde{u}_i$ is the amplitude of the i-th mode. It can be shown [43] that matrices $\mathbf{K}$, $\mathbf{M}$ and $\mathbf{C}$ are simultaneously diagonalized by

$$\begin{aligned} \mathbf{\Phi}^T\mathbf{M}\mathbf{\Phi} &= \mathbf{I} \\ \mathbf{\Phi}^T\mathbf{K}\mathbf{\Phi} &= \mathbf{\Omega}^2 \end{aligned} \tag{5.5}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{\Omega}^2$ is the diagonal matrix whose elements are the eigenvalues ω_i, $i = 1, ..., N$.

Premultiplying (5.1) by $\mathbf{\Phi}^T$ and substituting the displacement vector with its equivalent form in (5.2) leads to:

$$\ddot{\tilde{\mathbf{U}}} + \tilde{\mathbf{C}}\dot{\tilde{\mathbf{U}}} + \tilde{\mathbf{\Omega}}^2\tilde{\mathbf{U}} = \tilde{\mathbf{F}} \tag{5.6}$$

where $\tilde{\mathbf{C}} = \mathbf{\Phi}^T\mathbf{C}\mathbf{\Phi}$ and $\tilde{\mathbf{F}} = \mathbf{\Phi}^T\mathbf{F}$. The above matrix-form equation can be decoupled for each dimension into N scalar equations of the form:

$$\ddot{\tilde{u}}_i(t) + \tilde{c}_i\dot{\tilde{u}}_i(t) + \tilde{\omega}_i^2\tilde{u}_i(t) = \tilde{f}_i(t). \tag{5.7}$$

The solution of these equations at time t leads to the calculation of the amplitudes $\tilde{u}_i(t)$, $i = 1, ..., N$ and the deformation of the model is estimated using the modal superposition equation (5.4). At each time step, the new positions of the nodes of the model $\mathbf{X}(t)$ are given by

$$\mathbf{X}(t) = \mathbf{X}(t_0) + \mathbf{U}(t) \tag{5.8}$$

where $\mathbf{X}(t_0)$ is the vector containing the initial spatial positions of the model and $\mathbf{U}(t)$ is the nodal displacement vector.

In practice, the nodal displacements $\mathbf{U}(t)$ are approximated by $\hat{\mathbf{U}}(t)$ using a fraction of the modes of vibration, which present the highest amplitudes, that is:

$$\hat{\mathbf{U}}(t) = \sum_{i=1}^{l} \tilde{u}_i(t)\phi_i \tag{5.9}$$

where $l \ll N$. For the choice of the number of modes l, the total energy is calculated by:

$$E = \sum_{i=1}^{N} \tilde{u}_i^2 \tag{5.10}$$

and we chose the first l amplitudes carrying a predefined percentage of the total energy.

An issue that must be clarified is the calculation of the eigenvectors and eigenvalues of the generalized problem of (5.3). From the classical theory of vibration of a crystal lattice, it can be proved that the relationship between spatial (k) and temporal (ω) frequencies is given by:

$$\omega^2(p) = \frac{4K}{M} \sin^2\left(\frac{k(p)\alpha}{2}\right), \tag{5.11}$$

where α is the distance between the points of a closed chain. In (5.11), due to the periodicity of the closed chain:

$$k(p)\alpha = \frac{2\pi p}{N}, \ \ p \in \mathcal{B}(\mathcal{N}) \tag{5.12}$$

where $\mathcal{B}(\mathcal{N})$ is the Brillouin zone [84]:

$$\mathcal{B}(\mathcal{N}) = \begin{cases} \left[-\frac{N}{2}+1, ..., \frac{N}{2}\right], & \text{for } N \text{ even} \\ \left[-\frac{N-1}{2}, ..., \frac{N-1}{2}\right], & \text{for } N \text{ odd} \end{cases} \tag{5.13}$$

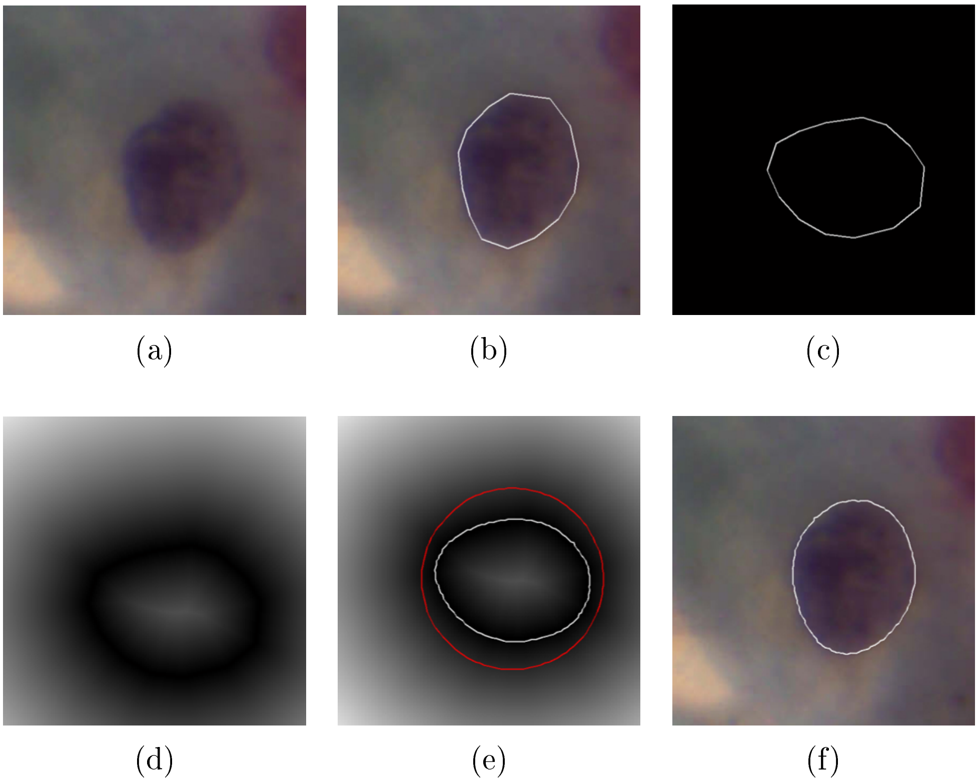
Combining (5.11) and (5.12) we can calculate the temporal frequencies ω_i^2, which correspond to the eigenvalues of the problem in (5.3). The corresponding eigenvectors $\phi(p)$ are then given by:

$$\phi(p) = \left[..., \cos\frac{2\pi p}{N}, ...\right]^T \tag{5.14}$$

Thus, using (5.14), analytic forms for the eigenvectors are obtained and the motion of the model can be easily expressed in terms of frequency modes as described in (5.9).

Training the physical model

Instead of describing the object of interest by a set of n labeled landmark points as in the standard ASM algorithm, we focus on the learning of the generalized displacements $\tilde{\mathbf{U}}$ of the model, in each image of the training set. This is an equivalent representation, since the combination of (5.2) and (5.8) results in the spatial coordinates of the shape. In the training phase, the nuclei boundaries were manually traced by an expert in all the images of the training set. An issue that must be taken into account for the correct training of the model is that the shapes in each image must be registered. Given the fact that the nuclei generally follow an ellipse-like shape, we registered all the manually traced shapes with a reference shape having its major axis horizontally oriented. Based on this boundary, the distance transform was estimated for every image. On the resulted image, a physical model was initialized and deformed until convergence, in order to detect the desired boundary. As a result, an accurate nucleus boundary was obtained. This procedure is depicted in figure 5.1(b).


(a) (b) (c)

(d) (e) (f)

Figure 5.1: Convergence of the physical model to the boundary of a cell nucleus in the training phase. (a) The initial image, (b) the manually traced nucleus boundary, (c) registration of the boundary to a reference shape oriented horizontally, (d) the distance transform of (c), (e) the initial (red) and the final (white) position of the physical model, (f) the detected nucleus boundary in the initial image rotated appropriately. The figure is better seen in colour.

From the final shape of the model, the generalized displacement vector $\tilde{\mathbf{U}}$ was estimated and from the entire training set, the mean $\overline{\tilde{\mathbf{U}}}$ was calculated, which entails in the representation of the mean shape of the nucleus boundary. More specifically, given a set of L vectors $\tilde{\mathbf{U}}_i$, the mean is calculated as:

$$\overline{\tilde{\mathbf{U}}} = \frac{1}{L} \sum_{i=1}^{L} \tilde{\mathbf{U}}_i \tag{5.15}$$

The covariance of the vectors is calculated by:

$$\mathbf{S} = \frac{1}{L-1} \sum_{i=1}^{L} \left(\tilde{\mathbf{U}}_i - \overline{\tilde{\mathbf{U}}} \right) \left(\tilde{\mathbf{U}}_i - \overline{\tilde{\mathbf{U}}} \right)^T \tag{5.16}$$

Using principal component analysis (PCA), the eigenvectors $\mathbf{a}_i$ with the corresponding eigenvalues λ_i of the covariance matrix $\mathbf{S}$ are used for an equivalent representation of the shape, that is

$$\tilde{\mathbf{U}} = \overline{\tilde{\mathbf{U}}} + \mathbf{A}\mathbf{b}, \tag{5.17}$$

where $\mathbf{A}$ is the matrix with columns the eigenvectors $\mathbf{a}_i$ and $\mathbf{b}$ is a vector containing the model coordinates in the basis of the eigenvectors:

$$\mathbf{b} = \mathbf{A}^T \left(\tilde{\mathbf{U}} - \overline{\tilde{\mathbf{U}}} \right) \tag{5.18}$$

Taking into account the J eigenvectors which correspond to the J largest eigenvalues of the covariance matrix, the shape can by approximated by:

$$\tilde{\mathbf{U}} \simeq \overline{\tilde{\mathbf{U}}} + \mathbf{A}_J \mathbf{b}_J = \left[\mathbf{a}_1 \, \mathbf{a}_2 \, ... \, \mathbf{a}_J \right] \left[b_1 \, b_2 \, ... \, b_J \right]^T \tag{5.19}$$

where $\mathbf{A}_J$ and $\mathbf{b}_J$ are derived from $\mathbf{A}$ and $\mathbf{b}$ by using only the J selected eigenvectors. Figure 5.2 shows the modes of variation of the shape of the nuclei along four individual eigenvectors $\mathbf{a}_j$, $j = 1, ..., 4$, extracted from the learning phase. Thus, using (5.15) and (5.9), the mean shape of the nuclei is described and this will be used as an initial template, in order to separate the overlapped nuclei in the images of the test set.

5.2.2 Segmentation of overlapping nuclei boundaries

This procedure includes the determination of the initial positions of the two models in the image and the deformation process that the models follow until convergence. A graphical description of this step is depicted in Figure 5.3 and the details are described in the following paragraphs.

Initialization of the model

The most important prerequisite of our method in order to provide reliable results is the accurate localization of the initial model. If the initial model is not close to the real boundary, then the results would probably be highly erroneous, as the model would converge in local minima of the image, which do not correspond to the real nucleus

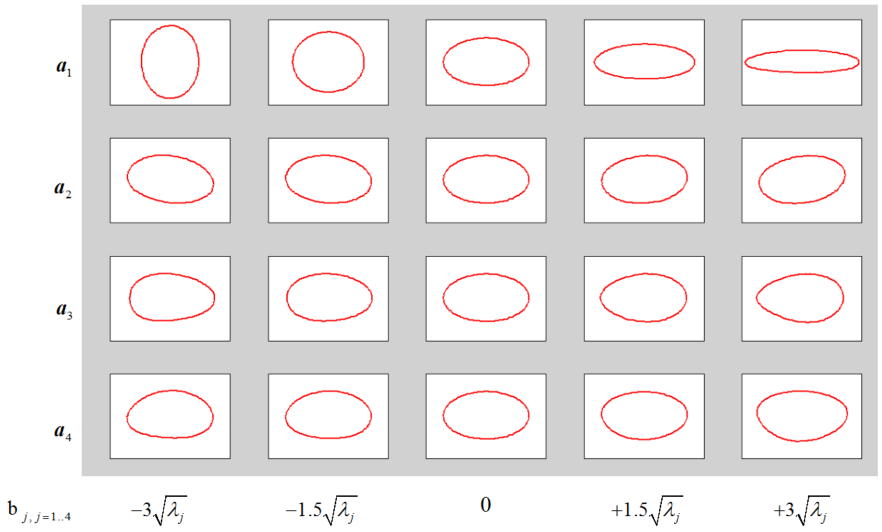


Figure 5.2: Modes of variation of the nuclei shape. Each row corresponds to the variations of the shape of the model by approximating $\tilde{\mathbf{U}}$ using only one eigenvector $\mathbf{a}_j$, $j = 1, ...4$ and setting the corresponded value of b_j to the values in the last row, with all the values of $b_k = 0$, for $k \neq j$ in (5.19). In the middle column, the mean shape is depicted.

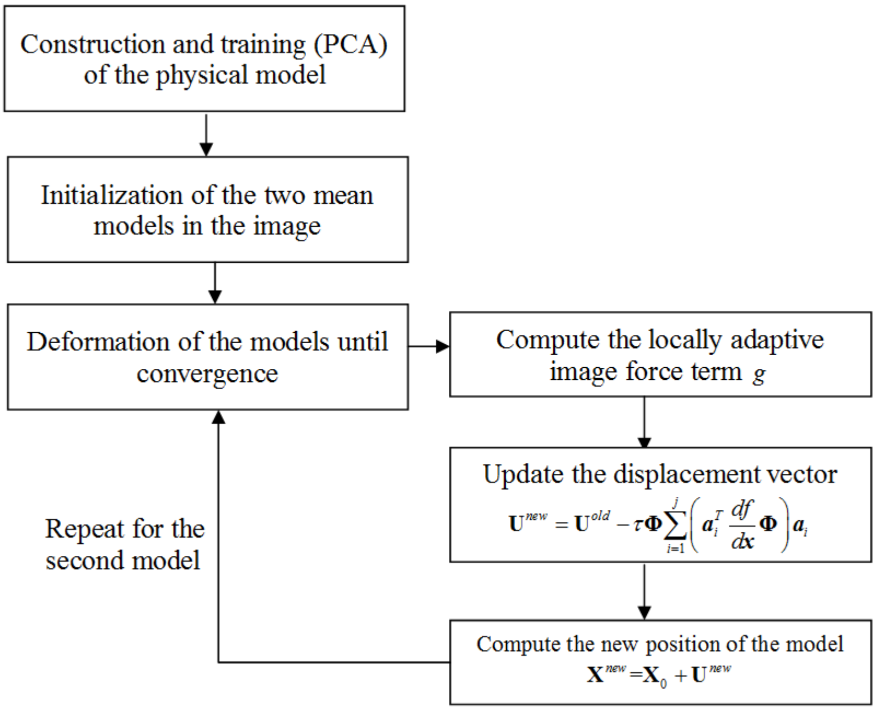

Construction and training (PCA) of the physical model
Initialization of the two mean models in the image
Deformation of the models until convergence
Compute the locally adaptive image force term g
Update the displacement vector
$\mathbf{U}^{new} = \mathbf{U}^{old} - \tau\mathbf{\Phi}\sum_{i=1}^{j}\left(\mathbf{a}_i^T \frac{df}{dx}\mathbf{\Phi}\right)\mathbf{a}_i$
Compute the new position of the model
$\mathbf{X}^{new} = \mathbf{X}_0 + \mathbf{U}^{new}$
Repeat for the second model

Figure 5.3: The basic steps of the segmentation of the overlapping nuclei (see text for details).

boundary. The mean nuclei boundary that has been determined in the training phase stands as the initial model of the nucleus boundary. As we are looking for two nuclei in the images of the test set, the initial positions of two models must be detected close to the real nuclei boundaries. For this reason, each image is first preprocessed for the detection of the strong nuclei edges, which will force the mean nuclei boundary to be located near to the real one through chamfer matching [92].

More specifically, in the grayscale counterpart of the initial image, we first reduce the noise by applying a Gaussian filter. Then, using histogram equalization technique, the contrast of the nuclei and the background is enhanced. In order to avoid the inhomogeneities inside the nuclei areas, which are commonly present due to uneven staining of the smear, we proceed with the formation of homogenous minima intensity valleys. This is feasible with the application of the H-minima transform in the original image, which is a grayscale morphological reconstruction [68]. Thus a marker image is constructed by the subtraction of a threshold value h from every pixel of the complement of the initial image. Then through the grayscale reconstruction process, we obtain an image that contains the regional minima, whose depth is less than h, suppressed. The result of this process is depicted in Figure 5.4(b), where a rough description of the positions of the two overlapped nuclei is defined. In this image, the Canny edge detector is applied (Figure 5.4(c)), and some strong boundary edges are detected. The distance transform g is then calculated [93].

To obtain the distance potential force, a distance surface is first built, that is:

$$g(x,y) = \min_{(p,q)\in\{(a,b):BW(a,b)=1\}} [d(x,y;p,q)] \tag{5.20}$$

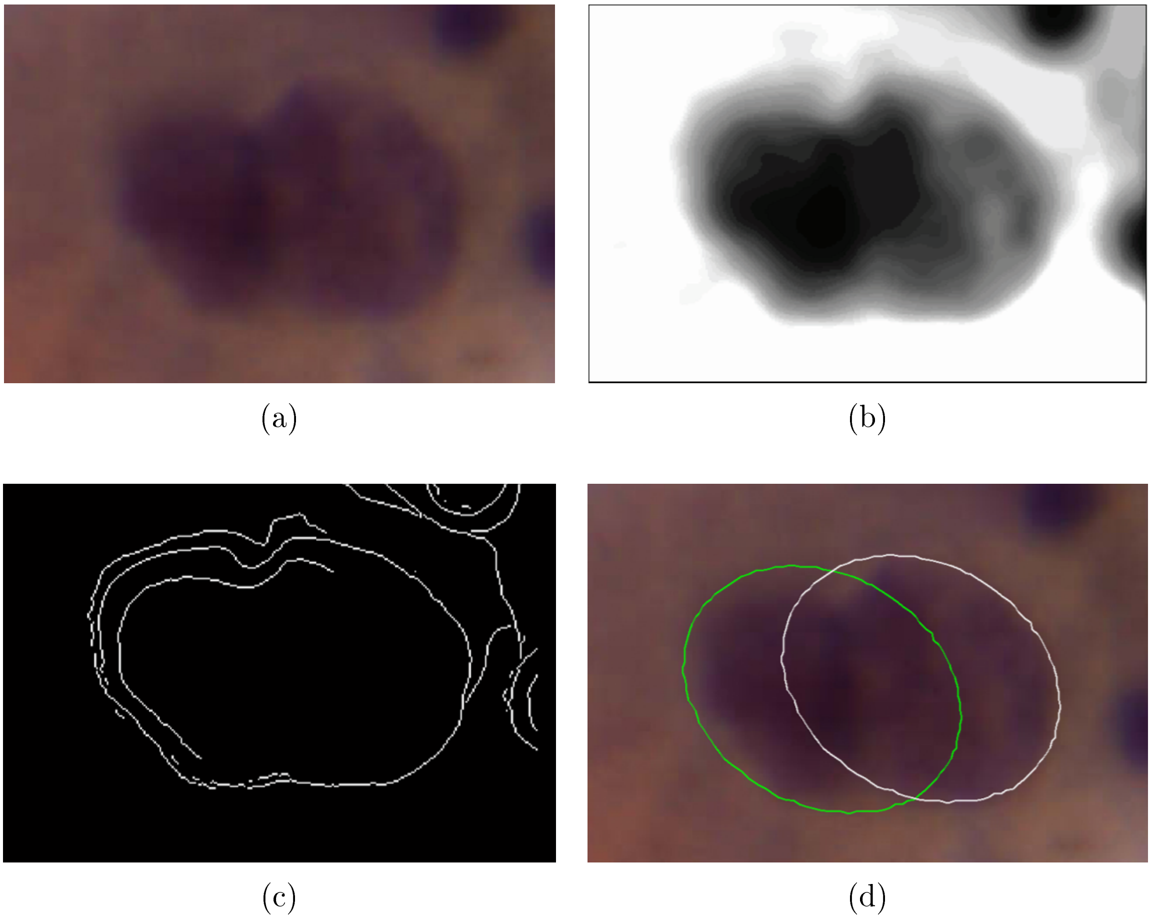
where $d(x,y;p,q)$ is the Euclidean distance between pixels (x,y) and (p,q) in the binary image BW, which is obtained after the application of the Canny edge detector. It must be noted that the location (p,q) corresponds to the location of a detected edge.

In the resulting image we search for the best two matching positions of the initial model (Figure 5.4(d)). The measure of correspondence between the edges and the model is the sum of the pixel values at which the model is located. A perfect match would produce a zero value in this measure (as the model would perfectly match in the edges of the image, in which the value of the distance transform is zero). However, as this is an extremely rare case in real images, we search for the position of the model in the image that minimizes this sum.

It must be noted that in each image the nuclei size and orientation may vary. Thus, the initial model is rotated by a step of 1° angle and scaled by factors between 0.6 and 1.2 of its original position and size. In this way, more accurate initial approximations of the nuclei boundaries are detected (Fig. 5.5). After the detection of the initial position of the models, we proceed with the deformation of the models in order to converge to the final nuclei boundaries.

Deformation of the models

Using the shape representation defined in (5.19), the algorithm fits the desired model in the image, driven by the image characteristics and the prior training. In each iteration,


(a)

(b)

(c)

(d)

Figure 5.4: (a) The initial image, (b) image obtained after noise reduction, histogram equalization and H-minima transform of the greyscale counterpart of the initial image, (c) the result of the application of the Canny edge detector in (b), (d) initial placement of the learnt model on each nucleus, after the chamfer matching. The figure is better seen in colour.

the changes in the generalized displacements should be consistent with the learnt parameters, and this is feasible by the minimization of a cost function $f\left(\tilde{\mathbf{U}}\right) = g\left(\mathbf{X}_0 + \boldsymbol{\Phi}\tilde{\mathbf{U}}\right)$, where g is the distance transform of the image as it is defined in the initialization step and its argument is the deformed shape with respect to (5.8) by omitting the temporal dependency for simplicity. More specifically, in each step, the algorithm selects the new generalized displacements by the following optimization schemes:

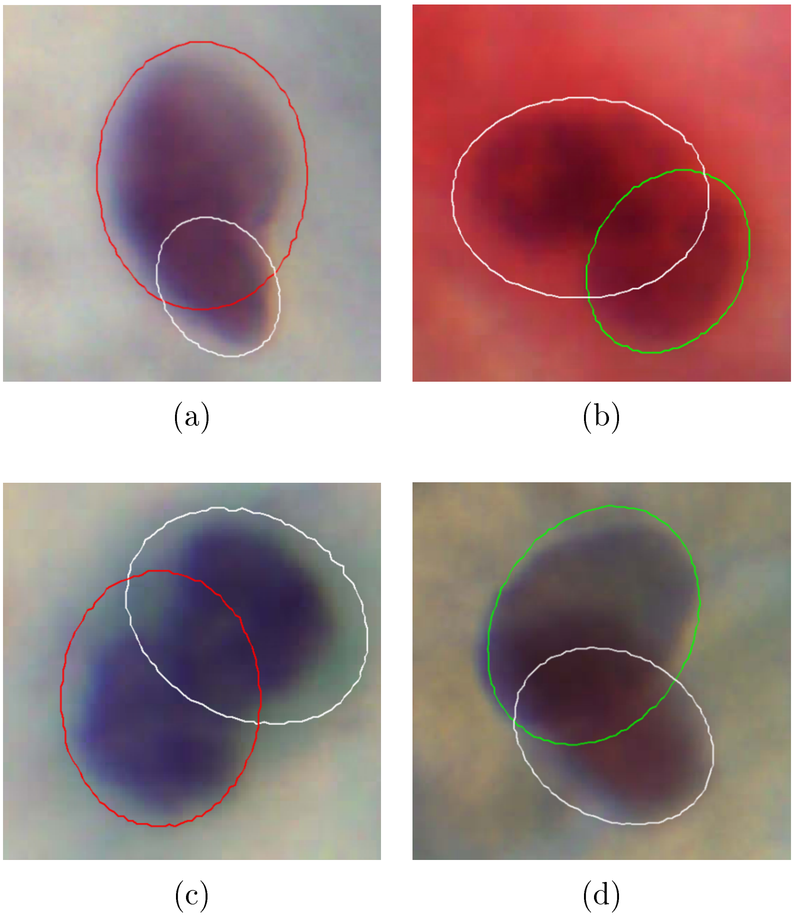
$$\min_{\tilde{\mathbf{U}}} f(\tilde{\mathbf{U}}) = \min_{b_i,\, i=1\ldots,J} f\left(\overline{\tilde{\mathbf{U}}} + [\mathbf{a}_1\ \mathbf{a}_2\ \ldots\ \mathbf{a}_J]\,[b_1\ b_2\ \ldots\ b_J]^T\right) \tag{5.21}$$

The gradient descent scheme of (5.21) with the new variables b_i is given by:

$$b_i^{\text{new}} = b_i^{\text{old}} - \tau\left(\mathbf{a}_i^T \frac{df}{d\tilde{\mathbf{U}}}\right) \tag{5.22}$$

where τ is the time step and:

$$\frac{df}{d\tilde{\mathbf{U}}} = \frac{df}{d\mathbf{X}}\frac{d\mathbf{X}}{d\tilde{\mathbf{U}}} = \frac{df}{d\mathbf{X}}\boldsymbol{\Phi} \tag{5.23}$$



(a) (b)

(c) (d)

Figure 5.5: Initial positions of the models in several images. Notice that best initial positions are obtained with the rotation and the scaling of the learnt model. The figure is better seen in colour.

The term

$$\frac{df}{d\mathbf{X}} = [\nabla g(x_1, y_1), ..., \nabla g(x_N, y_N)]^T \tag{5.24}$$

is actually the gradient of the image force term, where g is again the distance transform of the image as it is described above. In terms of the original variable $\tilde{\mathbf{U}}$, the update rule (5.22) turns out to be

$$\begin{aligned} \tilde{\mathbf{U}}^{\text{new}} &= \tilde{\mathbf{U}}^{\text{old}} - \tau \sum_{i=1}^{J} \left(\mathbf{a}_i^T \frac{df}{d\tilde{\mathbf{U}}} \right) \mathbf{a}_i \qquad (5.25) \\ &= \tilde{\mathbf{U}}^{\text{old}} - \tau \sum_{i=1}^{J} \left(\mathbf{a}_i^T \frac{df}{d\mathbf{X}} \boldsymbol{\Phi} \right) \mathbf{a}_i \end{aligned}$$

Notice that the initial value for $\tilde{\mathbf{U}}$ is the mean shape obtained with the training of ASM. Premultiplying (5.25) by $\boldsymbol{\Phi}$ to the left we get

$$\boldsymbol{\Phi}\tilde{\mathbf{U}}^{\text{new}} = \boldsymbol{\Phi}\tilde{\mathbf{U}}^{\text{old}} - \tau \boldsymbol{\Phi} \sum_{i=1}^{J} \left(\mathbf{a}_i^T \frac{df}{d\mathbf{X}} \boldsymbol{\Phi} \right) \mathbf{a}_i \tag{5.26}$$

Regarding (5.22), we have

$$\mathbf{U}^{\text{new}} = \mathbf{U}^{\text{old}} - \tau\mathbf{\Phi}\sum_{i=1}^{J}\left(\mathbf{a}_i^T\frac{df}{d\mathbf{X}}\mathbf{\Phi}\right)\mathbf{a}_i. \tag{5.27}$$

Finally, from (5.8), the local positions of the landmark points are calculated by:

$$\mathbf{X}^{\text{new}} = \mathbf{X}_0 + \mathbf{U}^{\text{new}}, \tag{5.28}$$

where $\mathbf{X}_0$ is the initial position of the mean shape of the model in the image. The models deform until convergence, i.e. no significant change is observed in the location of the points consisting the models between two sequential steps.

Spatially adaptive image force

The image force is defined as the force due to the potential field created by the image characteristics. The most common approach is to use the image gradient magnitude as the external force, in order to guide the deformable model in the areas of the image where high gradients are located (which usually imply the existence of strong edges of the objects of interest). However, the limitation of the gradient image force becomes evident in parts of the image with smooth intensity transitions, in which the gradient magnitude is very low. In Pap smear images, the staining procedure introduces variances in illumination and dye concentration. In some cases the nuclei borders may not be clearly distinguishable from the background, and these locations present weak image gradient. If a deformable model is initialized in such locations, it is not probable to guide it toward an edge. Thus, the gradient-based force field has a limited capture range for the deformable model. The distance potential force alleviates this issue for binary images.

As we can observe from (5.27) and (5.28), the deformation of the model is also controlled by the image force term. The degree of the influence of the image force term in the motion of the deformable model can be modulated by setting appropriate weight values $w_1, w_2, ..., w_n$ at each image point that belongs to the model, and the image force can be defined as

$$\frac{df}{d\mathbf{X}} = \left[w_1\nabla g(x_1, y_1), ..., w_N\nabla g(x_N, y_N)\right]^T. \tag{5.29}$$

Thus, if the weights of this term have large values, the model will deform mainly according to the image characteristics. On the other hand, if the image force weight is small, the model deformation would be driven by the learnt nuclei shape. In images of overlapping nuclei, the edges of the isolated part of the nuclei boundary must be taken into account, in order to attract the model for the detection of the true nuclei boundaries. In those points, high value of the image force term is desirable. However, in the area of overlap, there is no edge information and with the use of high weight values, the model could be attracted from the boundary of the isolated part of the neighbor nucleus, resulting in a high erroneous identification of the real nucleus boundary. An immoderate example of the influence of the values of the image force weights in the deformation of the models is depicted in Figure 5.6, where we attempt to emphasize the influence of different weight values in the final segmentation results. For this reason, we have used extreme weight values and we let the deformable model move without any restrictions, using (5.26) and (5.27). As we can observe, the segmentation results are highly erroneous.

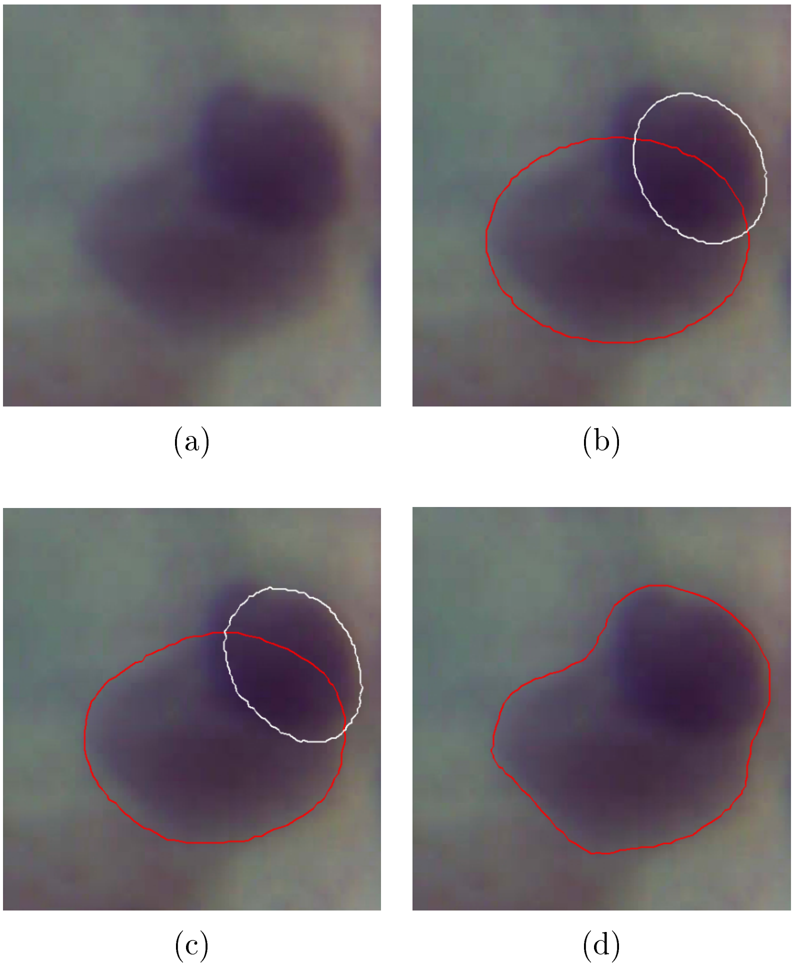


Figure 5.6: (a) Initial image and, (b) initial position of the two models. Result obtained using (c) small and (d) large weights values for the image force of one of the two nuclei. Notice that in (c) there exist small differences from the initial position of the models. However in (d), the model of one nucleus converges in a position of high gradient of the image, resulting in the erroneous identification of the boundary of the area containing both of the nuclei. The figure is better seen in colour.

In order to avoid such phenomena, we use different values for the weights in the image force term, depending on the position of the point in the boundary (Fig. 5.7). Thus, in each step of the deformation of the two models, the area of overlap is determined, and the points of the models lying in this area are associated with small weight values compared with the weight values in the non overlapping area, as it is explained in the next paragraphs. Therefore, the influence of the image force in the deformation of this part of the model is limited. However, in the rest of the points of the model, we use large values for the image term weights, and this results in the detection of the actual nuclei boundaries in the non-overlapping area of each nucleus. The PCA coordinates $\mathbf{a}_i$ estimated by the non-overlapping part of the model yield a set of modal amplitudes which determine the behavior of the rest of the points based on the learnt vibrations.

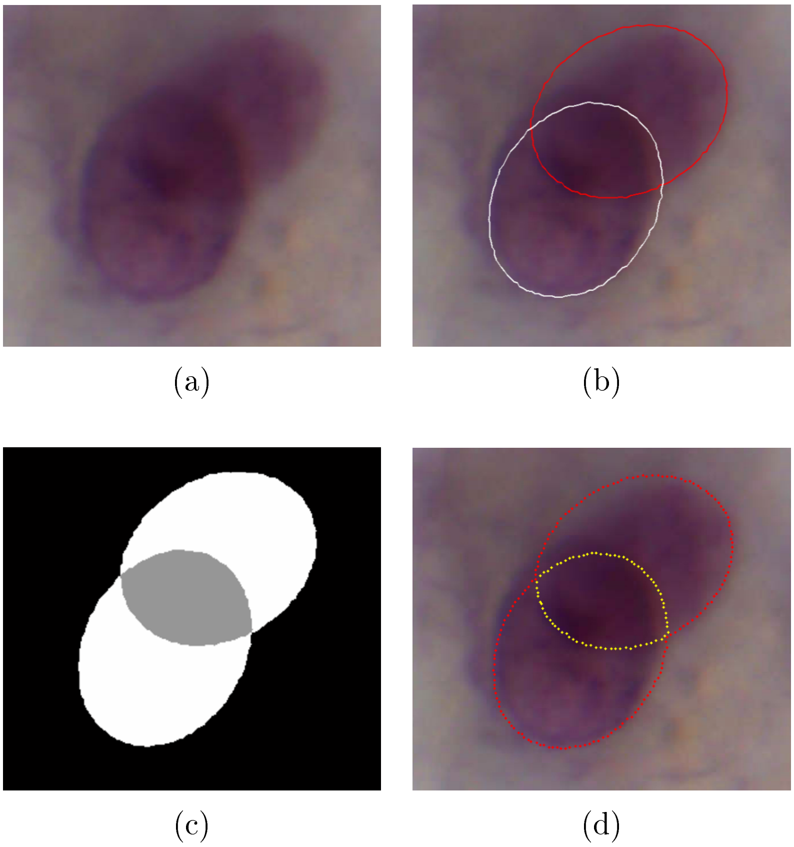


Figure 5.7: (a) Initial image, (b) initial position of the models, (c) calculation of the overlapping area, (d) points lying in the area of overlap, denoted with yellow, have very small weights with respect to the rest of the points. The figure is better seen in colour.

5.3 Experimental results and discussion

5.3.1 Study group

In our experiments, we have used 46 images containing a single normal nucleus of conventional Pap smear slides, in order to construct the training set used for the training of the physically based model. The proposed method was tested in terms of the accurate determination of the nuclei boundary on a test set of 50 images containing two overlapping nuclei each (as in Figure 5.4), yielding 100 nuclei in total. Thus, the training and the test set of images are independent. All of these images were acquired through a CCD camera (Olympus DP71) adapted to an optical microscope (Olympus BX51). We have used a

40× magnification lens and the acquired images were stored in JPEG format. The initial images obtained by the optical microscope have size 3072 × 4080 pixels. These images are then segmented manually in order to construct the database of images containing the overlapped nuclei and a small portion of the neighboring background (such as in Figures 5.4, 5.6, 5.10 and 5.11). The average size of all the images in our database is 260 × 300 pixels.

5.3.2 Numerical evaluation

For the evaluation of the performance of the method, the boundaries of the nuclei in the entire test set were estimated manually, after careful examination of an expert cytopathologist. The determination of the boundary of each nucleus in the overlapping area was based on the exploitation of small variances in the intensity of this area, which an expert could identify. However, in many cases, there was no existence of intensity variances, and the boundary of each nucleus was manually drawn by the expert, following the expected shape of the nucleus in the specific image. Based on the ground truth, we have calculated the mean (μ) and the standard deviation (σ) for the area of overlap in our data set. The overlapping area varies between 4.0% and 48.2% with 18.52 ± 11.77 (%) ($\mu \pm \sigma$).

In order to evaluate the performance of the proposed method, the Euclidean and Hausdorff distance of the final position of the model and the ground truth was calculated in each image. The presented method underwent a twofold evaluation: a) we have tested the influence of the use of different weight values in the final segmentation compared to the use of one single value and b) the results of the method were compared to the results of different segmentation methods, such as the standard ASM algorithm using different weight values in the area of overlap (as it is proposed in our method), the conventional technique of the ellipse fitting in the overlapped nuclei boundary incorporated in our method (instead of using different image force weight values), the unsupervised segmentation of overlapped nuclei using Bayesian classification [5] and the H-minima transform-based marker extraction and contour parameterization method for segmenting overlapped nuclei [4]. It must be noted that the ellipse fitting technique is extensively used from several researchers ([5, 4, 81]), in order to estimate the nucleus boundary in the overlapping areas and it is based on the hypothesis that the nuclei shape is generally ellipse-like. Thus, in each step of the deformation of the model, the points of the boundary of the nucleus of the non-overlapping area are used for the calculation of the interpolated ellipse using the direct least squares fitting of ellipses [94]. Then, the part of the boundary in the overlapping area is completed using the corresponding part of the interpolated ellipse and the link points are smoothed (they are calculated as the average of the previous and the next point of the model).

The choice of the weights in our work is performed with a two step procedure: First, we use the same value for the weights of overlapping and non-overlapping areas, and we test the method for several different values for the weights. The Hausdorff distance between the result of the proposed method and the ground truth using a single value for the image force term is depicted in 5.8. As we can see, the best results were obtained for $w_{1,\ldots,n} = 0.1$ and the Hausdorff distance for this weight is 24.29. However, with the use of different weight values, the corresponding Hausdorff distance is reduced, which indicates that the use of different weight values renders the method more performing. Thus, in the next step, having found that the best segmentation was obtained with the weight value 0.1 (Figure 5.8), we keep this value constant for the external weights in non overlapping areas and we test multiple values for the weights in the overlapping areas. This is described in Figure 5.9, which contains the segmentation results of the method from several experiments, where we have used different weight values for the overlapping areas, while the weight value in non overlapping areas was fixed to 0.1. From this image, we can observe that the best segmentation results were obtained for $w_{1,\ldots,k} = 0.001$, where k is the number of the points of the model lying in the overlapping areas.

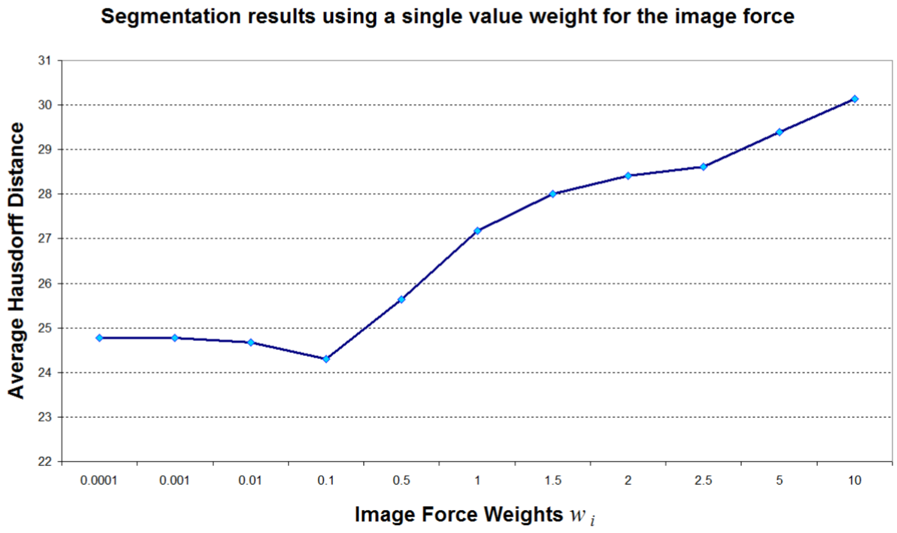


Figure 5.8: The result of the proposed method using the same value for the image force term, for all of the points of the model. Notice that best results are obtained using the value of 0.1 and the corresponding Hausdorff distance is 24.29.

The comparative results of the proposed method and the other segmentation techniques are included in Table 5.1. As we can see, our method exhibits better performance, since both the Euclidean and Hausdroff distances are smaller compared with the other methods, and it does not exhibit large variations, as the standard deviation is small. This implies that the proposed model is closer to the manually traced nuclei boundaries, and as a result it is more accurate than the other approaches. Furthermore, based on the experimental results and the overlapping percentage of our image data set (as it was described in the above paragraphs), we can conclude that our method can successfully segment nuclei that present grade of overlap lower than 50%.

Several examples of the segmentation results of our method are depicted in Figure 5.10. The segmentation results in images of 5.11. of our method and several methods proposed in the literature are depicted in 5.12. As we can see, the use of the ellipse fitting algorithm

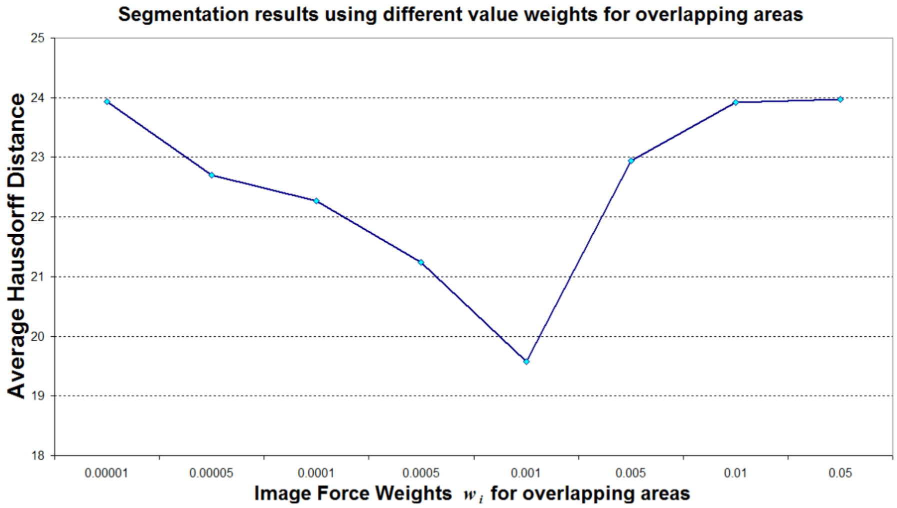


Figure 5.9: The result of the proposed method using the same value $w_i = 0.1$ for the image force term in the points lying in the non-overlapping area of the nuclei, and multiple values $w_{1,2,\ldots,k}$ for the k points of the model lying in the overlapping area. Notice that best results are obtained using the value of 0.001 and the corresponding Hausdorff distance is 19.58.

Table 5.1: Results of overlapped nuclei segmentation methods in terms of Euclidean and Hausdorff Distances ($\mu \pm \sigma$)

.

	Hausdorff	Euclidean
Proposed method	19.58 ± 8.52	8.64 ± 3.77
Ellipse fitting	20.73 ± 10.38	8.73 ± 3.90
ASM	21.33 ± 11.48	9.78 ± 4.10
Bayesian classification	24.18 ± 12.61	10.95 ± 4.79
H-minima	22.70 ± 11.90	10.71 ± 4.35

for the overlapped parts of the nuclei in our method produces some rugged points in the nuclei boundaries, and the boundaries of the overlapping areas are not well detected, as they do not follow an elliptical form in the overlapping areas. Furthermore, the standard ASM algorithm produces noisy results and this indicates that the representation of the shape of the nuclei by the vibrations of a spring-mass system instead of landmark points, assures a smooth representation of the shape, and as a result more accurate boundaries for the nuclei.

It must be noted that in our work, we have trained the ASM having as training shapes the nuclei boundaries obtained with the convergence of the physical model in the training set, and not with independently distributed points as it is the standard procedure. By these means, the ASM method is unbiased with respect to the manual tracing of the nuclei and the comparison of the two methods may be performed on the same basis. Besides,

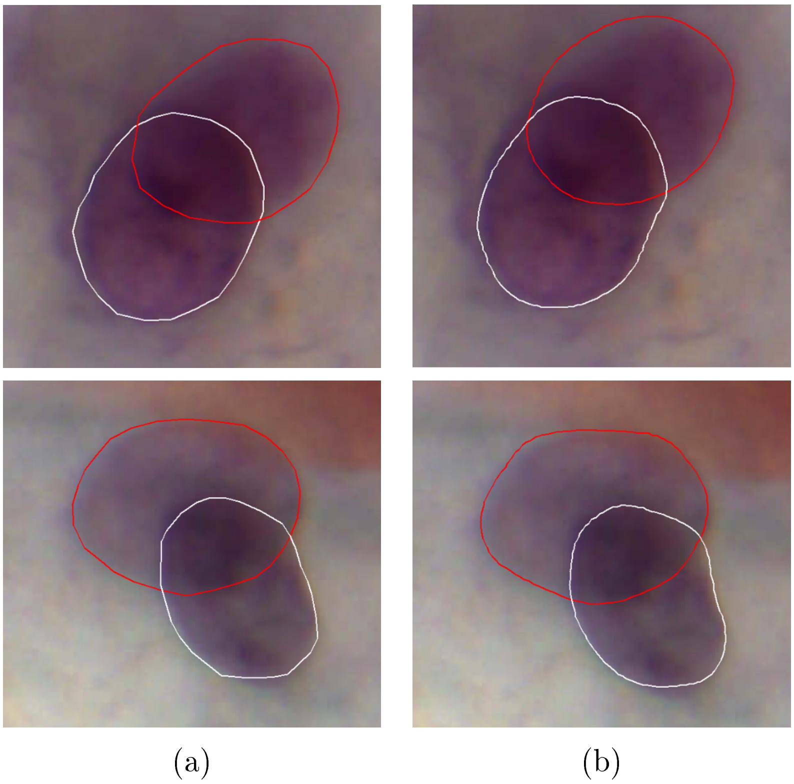


Figure 5.10: Segmentation results of the method. (a) Ground truth, (b) nuclei boundaries obtained with the convergence of the models. The figure is better seen in colour.

these shapes can be considered as an accurate representation of the nuclei boundaries, since the mean Euclidean distance of the final position of the physical model and the ground truth is 1.76 pixels in the images of the training set. For this reason, the main eigenvectors for the ASM model do not exhibit significant differences from the main eigenvectors of the physical model. However, the noisy results of the ASM method are due to the minimization procedure, in which the ASM is more sensitive to the image force. This can be observed in Figure 5.12(d), where the discontinuities in the detected edges (Figure 5.11(b), second image, points A and B) produce rugged segments of points in the convergence of ASM. Furthermore, the detected edges in the inner area of the nucleus (Figure 5.11(b), first image, point A) seem to interfere to the convergence of the model, which fails to identify the actual nuclei borders. The reason for this effect is that in the standard ASM algorithm, the image force term $\frac{df}{d\mathbf{X}}$ has a direct influence on the new variables (the projection of the points of the model on the eigenvector space), which are then multiplied by the eigenvectors to obtain the new position of the model. On the other hand, in our case, the new variables (the projection of the generalized displacements on the eigenvector space) change according to $\frac{df}{d\tilde{\mathbf{U}}}$ in (5.22), or equivalently according to $\frac{df}{d\mathbf{X}}\mathbf{\Phi}$ in (5.27). Thus, the new position of the model depends on the specific product, which provides a smooth image force term. In this way, the model is not affected by the potential noise contained in the image force.

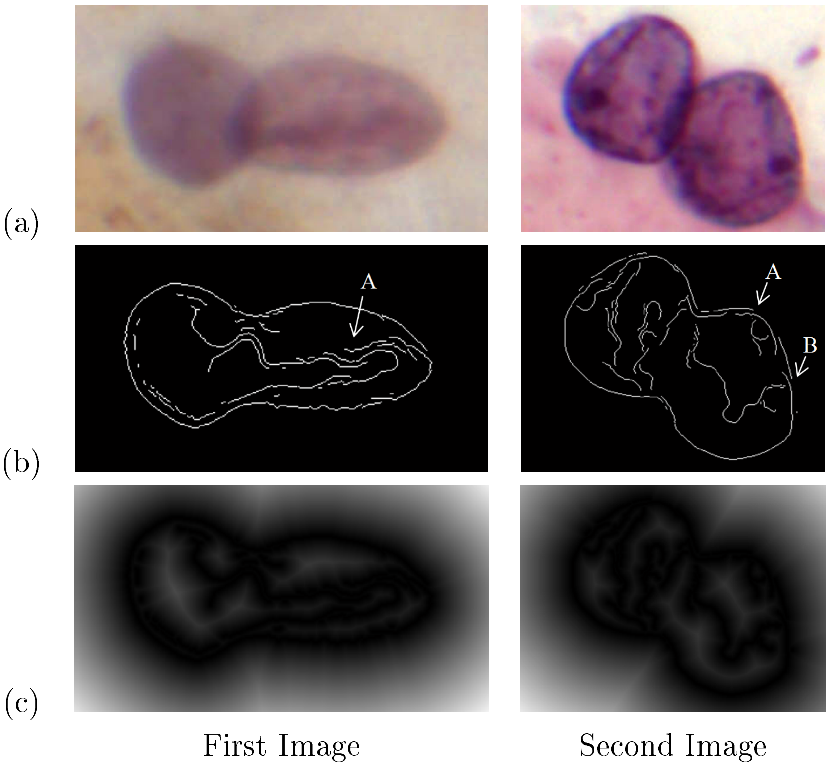


Figure 5.11: (a) Representative images containing overlapped nuclei. (b) Edge images resulted after the application of the Canny edge detector (initialization step). The point A in the first image indicates some falsely detected edges in the inner area of the nucleus. The points A and B in the second image indicate the existence of discontinuities. (c) The corresponding distance transform.

For the methods proposed in [5] and [4] the results are not quite accurate, and this is a consequence of several reasons. In [4] the outcome is actually the ellipse obtained by the detected nuclei boundary points produced by adaptive thresholding. As we can observe in Figure 5.12(e), the estimated ellipses provide a rough approximation of the actual nuclei boundaries and additional processing is required for the detection of more refined nuclei borders. The method in [4] may be considered similar to the initialization step of our algorithm; however, there are several differences with our method. More specifically, in the initialization step, we use the H-minima transform for the construction of smooth intensity valleys and the reduction of the noise in the image, in contrast with [4], in which the H-minima transform is used in a marker extraction scheme. For this reason we have used only one threshold value for h (Table 5.2) and we do not test many threshold values as it is proposed in [4]. Furthermore, the edges of the image are detected with the Canny edge detector, in order to obtain a binary image, without the use of global thresholding or the watershed transform as in [4]. In the binary image we try all possible similarity transformations, in order to achieve a reliable initial position of our model, which is known through the training step.

Furthermore, in [5] the boundaries of the non-overlapping area of the nuclei are ob-

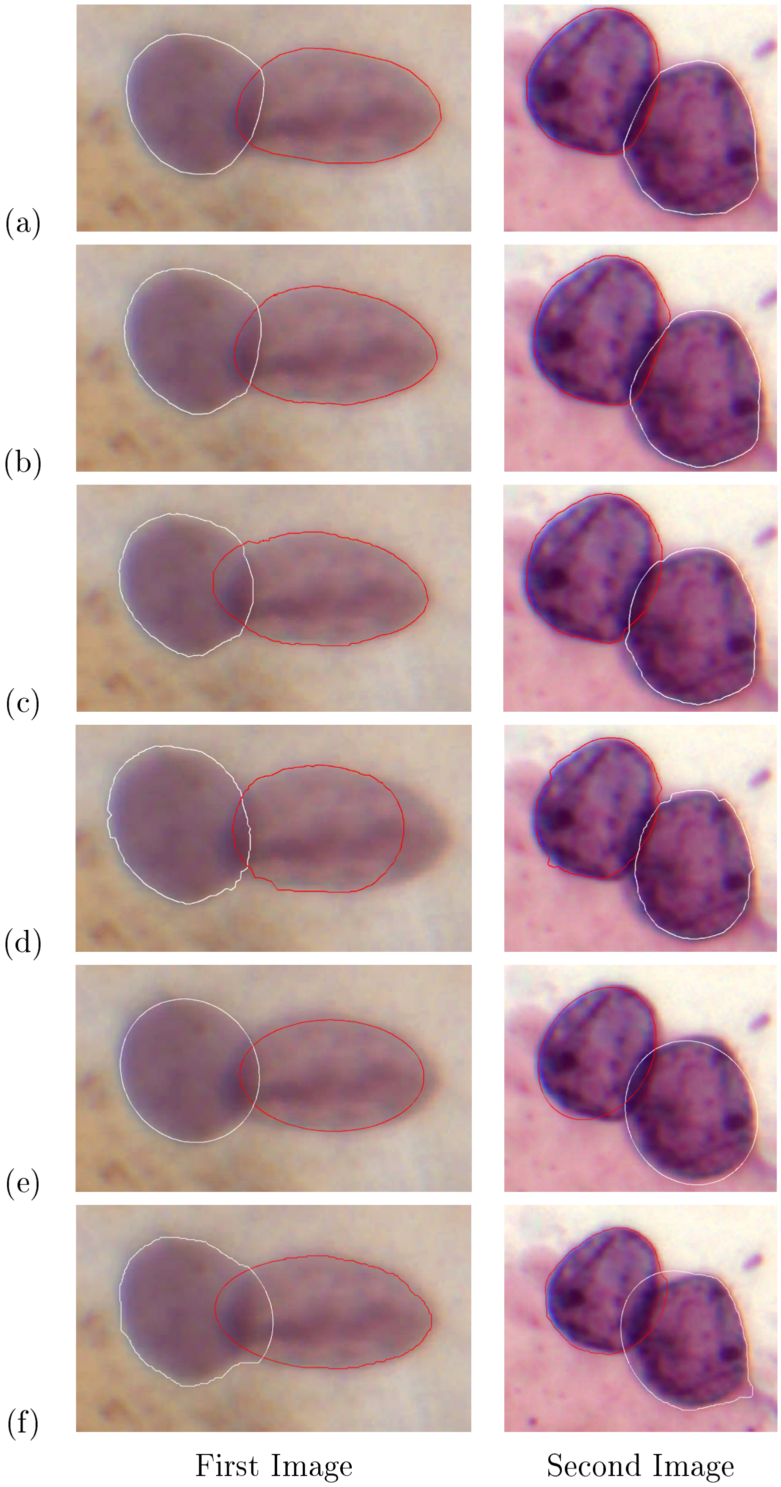


Figure 5.12: Segmentation results. (a) The ground truth in the initial image, the segmentation results for (b) the proposed method, (c) the ellipse fitting algorithm, (d) the standard ASM algorithm, (e) H-minima marker extraction [4], (f) Bayesian [5].

tained after morphological operations and adaptive thresholding. Although this procedure is fast, it does not always succeed in the detection of accurate boundaries. As we can

Table 5.2: Parameter values used in the experiments.

Parameter	Value
h (H-minima transform)	20
Stiffness K	5
Mass M	1
Dumping $\tilde{c}_i$, $i = 1, ..., N$	1
Radius of the initial model	85
Model points	120
Initial time step τ_0	0.001
Weights of image force (overlapping area)	0.001
Weights of image force (non overlapping area)	0.1
Number of modes $\tilde{u}_i$	14
Number of eigenvectors $\mathbf{a}_i$	4

see in Figure 5.12(f) (first image, white traced nucleus), the nucleus boundary that was erroneously detected in the bottom of the non-overlapping area, results in the estimation of an ellipse for the overlapping area that does not correspond to the expected nuclei boundaries. Furthermore, in Figure 5.12(f), (second image, white traced nucleus) we can observe that the inhomogeneity in the image intensity results in the wrong detection of the nucleus boundary in the non-overlapping area (near the bottom, right hand corner of the image). Therefore, we may conclude that in images with noise and artifacts (like Pap smear images), adaptive thresholding techniques do not provide accurate detection even for the non-overlapping boundaries of the nuclei and further processing is required.

Moreover, as a measure of the computational efficiency of the segmentation method, we present in Table 5.3 the mean times for the processing of each image (including the initialization and the deformation of the two models), developed in Matlab using a Pentium 2.0 GHz with 3GB RAM. As we can observe, the computational burden that the ellipse fitting algorithm introduces in each step of the deformation of the model, results in an increase of the processing time of each image. Furthermore, the reduction of the learnt parameters, as it is described in (5.19), renders our method superior to the standard ASM algorithm. In the cases of non iterative procedures, in which an ellipse fitting algorithm is implemented such as in [5] and [4] the required processing times are clearly shorter than in the rest of the methods. In both of these methods, the estimation of a suitable ellipse that fits the points in the non overlapping area is calculated once, and no iteration is performed. In contrast, the computational cost of the ellipse fitting algorithm in our method is bigger, because the ellipse is calculated in every iteration of the algorithm. The iterative procedure that we propose for the convergence of the model, although it requires more computational time, is necessary for the refinement of the nuclei borders in both the overlapping and non-overlapping areas, which is not achieved by the non-iterative methods. This can be easily verified by the comparative results in Table 5.1, where we can observe that the non-iterative methods have lower performance than the proposed method.

Table 5.3: Execution time ($\mu \pm \sigma$).

Segmentation method	Time (sec)
Proposed method	17.89 ± 0.79
Standard ASM	26.42 ± 2.90
Proposed method with ellipse fitting algorithm	46.21 ± 3.57
Bayesian classification [5]	4.41 ± 1.26
H-minima [4]	2.86 ± 0.64

The parameters of the steps of the method were selected after several tests (Table 5.2). Thus, the threshold h in the H-minima transform in the initialization of the models is set to 20. For the calculation of modal vibrations in (5.11), we selected $K = 5$ and $M = 1$. Furthermore, in (5.7), the value for all $\tilde{c}_i$, $i = 1, ..., N$ was set to 1. For the initialization of the physical model, a circle of radius 85 pixels and 120 points in its circumference was used centered at the centroid of the registered manually traced shape.

It must be noted that the change of the weight values for a large number of points in a single step may lead to a large displacement of the deformable model, as the additive value of the image force term in (5.27) would be large. In our implementation we have eliminated this effect with a variable time step, which is calculated in each iteration. Thus, given an initial time step $\tau_0 = 0.001$, the time step in (5.27) is calculated as $\tau_t = \tau_0(1 - OPR)$, where OPR is the overlapping points ratio, which is defined as the percentage of the points of the physical model lying in the overlapping area. By these means, a weighted sum of the contribution of the points in the non-overlapping and in the overlapping area is obtained which ensures the stability of the algorithm, as the deformation of the model is smooth and it avoids abrupt changes.

From (5.9), we have calculated that the first 14 modes $\tilde{u}_i$ contain more than 99% of the total energy in each image of the training set. Thus, only 14 parameters (instead of 120 landmark points) are sufficient for the accurate representation of the desired shape. Furthermore, after the application of PCA in these learnt parameters, only 4 eigenvectors which correspond to the highest eigenvalues lead to an almost exact shape representation, as they represent the 99.9% of the total energy. This clarifies that the proposed segmentation method provides a more compact shape representation which results in the reduction of the parameters to be learnt.

It must be noted that our method could be extended to be able to segment three overlapped nuclei as the main steps remain the same and the models deform independently. The only differences would be in the initialization of the three models in the image and the detection of the overlapping area. In the first case, the process which is described in section 5.2.2 could be followed, but instead of two models, we search for three models through the chamfer matching. In the second case, for each model, the area of overlap could be defined as the union of the areas of the two other models and the model under consideration. Thus, the weights of the specific model would be defined accordingly. In the case where the images contain unknown number of nuclei, then an initial step of counting the existed nuclei is necessary, such as in [81].

5.3.3 Relation between the weight values and the overlapping percentage

In order to investigate the influence of the selected weight values compared to the degree of overlapping, we have performed an experiment on a synthetic image, constructed from two individual images of the training set, each one containing a singe nucleus (Fig. 5.13). In these images, we have selected the area of each nucleus based on the ground truth. Then, in a new image, the areas of the nuclei were incrementally overlapped, and we have compared the segmentation results of our method with respect to the degree of overlapping of the nucleus area and the selected value weights. To this end, we have calculated the Hausdorff distance of the obtained boundary of one nucleus of interest, whose area overlapping was ranging from 10% to 90% (Figure 5.14). The gray level value of the area of overlap of the examined nucleus is set to 95% of the initial gray level. The rest of the nuclei areas contained the same intensity value as in the original images and the background was set to the mean value of both images. It must be noted that, in the synthetic image, we considered that the initialization step produced acceptable initial positions for the models.

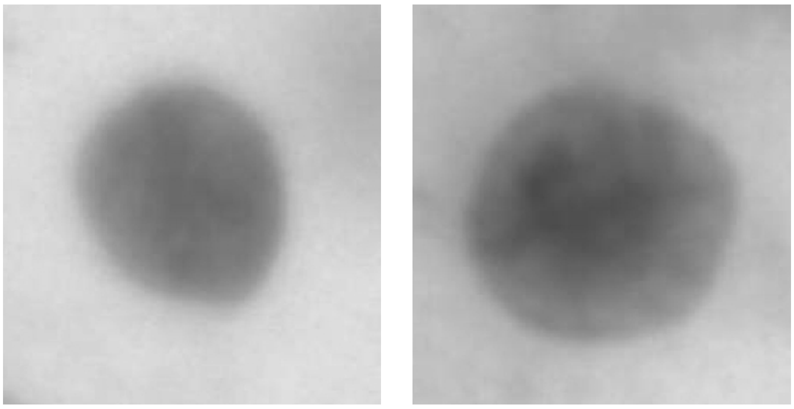


Figure 5.13: The original images of the training set used for the construction of the synthetic image.

The method was applied in this image for the drawing of conclusions about the influence of different rates between the weight values of overlapping and non overlapping areas, and how they affect the segmentation results with respect to the degree of overlap. Thus, the weight value for the points lying in the non-overlapping area was fixed to 0.1 for all the images. Then, the method was applied using different values for the weights of the

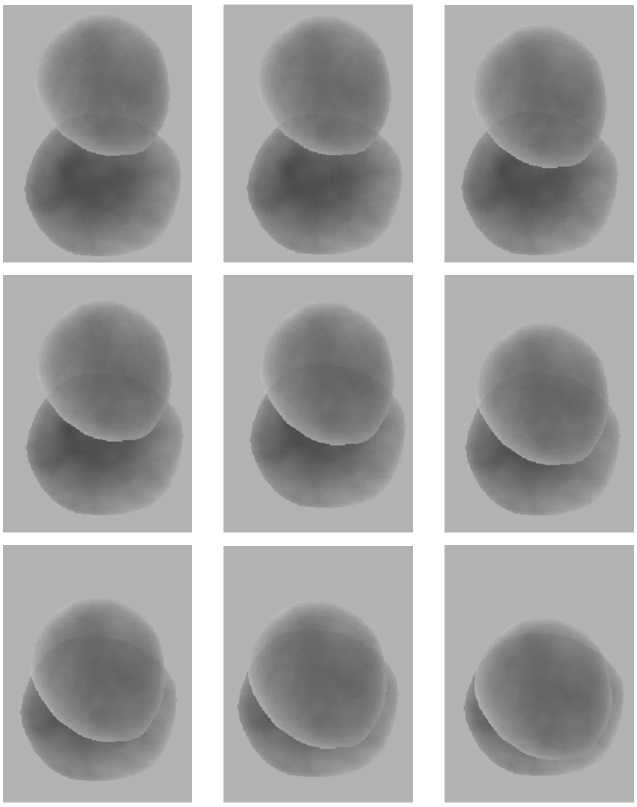


Figure 5.14: Construction of a synthetic image. The degree of area overlapping for the upper left nucleus is 10%, and for the bottom right is 90%.

nodes belonging to the overlapping areas. These weights were set to 2, 5, 10, 20, 100, 200 and 1000 times lower than the value of the weights of the nodes in the non overlapping areas. Furthermore an experiment with equal weight values of the overlapping and non-overlapping areas (set to 0.1) was also performed. The degree of overlap of the nucleus of interest in the synthetic images varied from 10% to 90%. As we can see from Figure 5.15, the choice of equal value for the weights produces worse results for almost all the cases of overlapping. Furthermore, values lower than 100 times of the non-overlapping weight value have the same result in the Hausdorff distance. In general, we can conclude that for overlapping percentage greater than 45% the best results were obtained using a weight value for the overlapping areas which is half than the weight value for the non overlapping areas. For smaller overlapping rates, the best results are obtained using a weight value of overlapping areas which is at least five times smaller than the weight value for the non overlapping areas. It must be noted, that the Hausdorff distance for this experiment is generally small, as the image of overlapping nuclei is, in a way, ideal, since it does not contain any noise or background artifacts.

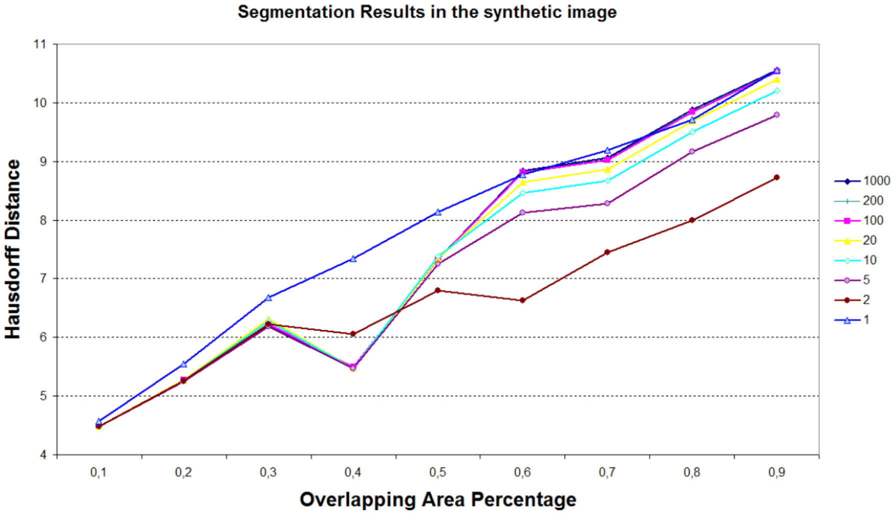


Figure 5.15: The Hausdorff distance in the synthetic image with respect to the nuclei overlapping percentage. The curves correspond to the segmentation obtained using the specific rates of the weight values of the non-overlapping areas to the weight values of the overlapping areas.

5.4 Conclusion

We have developed a segmentation method combining the physically based model, which provides a compact representation of the shape of the object of interest, and the active shape model, which takes advantage of the *a-priori* knowledge of the expected shape. The introduction of variable weights in the contribution of the image force in the deformation of the model results in the correct identification of the part of the nuclei boundary that lies in the overlapping areas. The method has been tested in terms of the accurate segmentation of the nuclei borders in images from Pap smear slides, and as it was verified by the results it presents a high performance. Thus, the method produces more accurate nuclei boundaries which are closer to the ground truth, compared to the standard ASM algorithm and the segmentation obtained by two methods proposed for the segmentation of overlapped nuclei. The main advantage of the proposed method is that it provides a flexible way for the simultaneous recognition of the isolated and the overlapped nucleus boundary. This avoids the development of an additional algorithm for the detection of the nuclei boundaries in occluded areas, such as the ellipse fitting algorithm.

Chapter 6

Cervical Cell Classification Based Exclusively On Nucleus Features

6.1 Introduction

The interpretation of Pap smear images relies basically on the visual recognition of the changes of the structural parts of the cells (nucleus and cytoplasm). However, this process is a tedious, time-consuming and in many cases error-prone procedure due to the high degree of complexity that these images exhibit. Several approaches have been proposed for the classification of cells in Pap smear images and they concern techniques such as Bayesian classifiers [55], artificial neural networks [56], support vector machines (SVM) [59] and nearest neighbor based classifiers [51]. It must be noted that most of these methods use presegmented images which contain only one cell, so the correct segmentation of the nucleus and the cytoplasm is feasible (Fig. 6.1(a)). In images containing cell clusters (Fig. 6.1(b)), the detection of the cytoplasm boundary is a difficult problem and until now, there is not any method in the literature that results in the automated delineation of the cytoplasm areas in cell clusters. However, the detection and segmentation of the nuclei in images containing cell overlapping and cell clusters has been successfully addressed by several studies [76, 77].

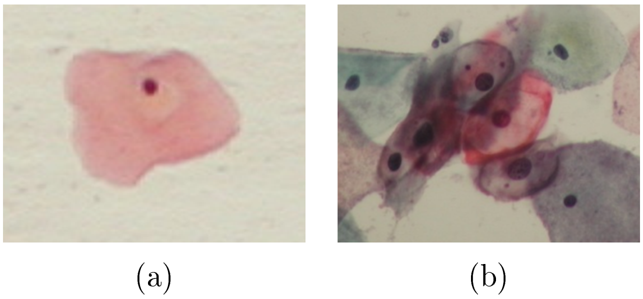


Figure 6.1: (a) An isolated cell and (b) overlapping cells. Notice that the cytoplasm area is clearly recognized in (a) while in (b) its determination is very ambiguous for each cell.

The methods which deal with the classification of Pap smear images are based on the calculation of features extracted from the areas of the nucleus and the cytoplasm [51, 52]. These features are usually based on shape and intensity characteristics of the objects of interest. However, the calculated features do not exhibit the same discriminative ability. For the determination of the most efficient feature set which will be used as input in a classifier, some feature selection schemes have been proposed, and they concern particle swarm optimization [52] and genetic algorithms [51].

Based on the aforementioned facts, we can conclude that there are two open problems in the automated classification of a Pap smear image acquired directly from an optical microscope: a) the limitation to use only the features extracted from the nuclei areas, as these are the areas that can be automatically segmented, and b) the determination of the most efficient feature subset, which will provide the best discriminative ability.

In this chapter, we evaluate the classification of cervical cells, based exclusively on nucleus features and ignoring the features extracted from the cytoplasm area. This is a crucial step for the correct characterization of Pap smear images acquired directly from an optical microscope, where the cell overlapping is an often found phenomenon and the delineation of the cytoplasm area can not be obtained automatically. In this direction, we investigate the representation of the features in low dimensional spaces using non linear dimensionality reduction methods. These techniques are advantageous in comparison with their linear counterparts, because they can properly handle complex nonlinear data, as they better describe the manifold where the data lie.

The low dimensional feature subsets serve as input in two unsupervised classifiers (spectral clustering [74] and fuzzy C-means [73]). As it was verified by the results, the non-linear dimensionality reduction techniques lead to a construction of nucleus-only feature subsets which can be used for the separation of normal and abnormal cells by the classifiers, presenting high performance.

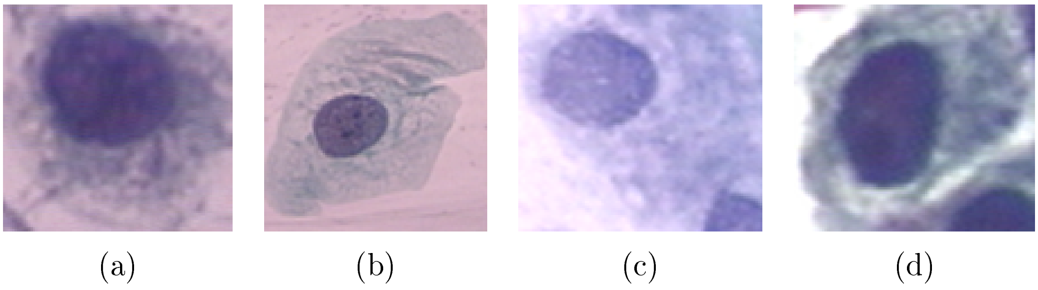


Figure 6.2: Types of cells included in the Pap smear benchmark [6]. (a)-(b) Abnormal cells and (c)-(d) normal cells.

6.2 Methodology

6.2.1 Study group

Our experiments are based on the Pap-smear benchmark database presented in [6] . The database consists of 917 images containing a single cell each (Fig. 6.2), and the samples are distributed unevenly in seven classes. Three of them are considered as normal and four of them are considered as abnormal types of cell. The detailed description of the database is depicted in Table 6.1.

Table 6.1: Distribution of cells in the Pap-smear benchmark database [6].

NORMAL	#cells
Superficial squamous epithelial	74
Intermediate squamous epithelial	70
Columnar epithelial	98
TOTAL	**242**
ABNORMAL	#cells
Mild squamous non-keratinizing dysplasia	182
Moderate squamous non-keratinizing dysplasia	146
Severe squamous non-keratinizing dysplasia	197
Squamous cell carcinoma in situ intermediate	150
TOTAL	**675**

6.2.2 Feature generation and dimensionality reduction

The images of the database have been manually segmented by experts and the areas of the nucleus and the cytoplasm are accurately defined. From these areas, twenty features concerning the intensity and the shape characteristics of the specific area are determined (Table 6.2). Nine out of twenty features concern the nucleus area and they can be calculated independently.

The techniques that we have used for the construction of the new feature sets concern non-linear dimensionality reduction schemes. In our study we have investigated the performance of four nonlinear techniques: Kernel-PCA [95], Isomap [96], Locally Linear Embedding [97] and Laplacian Eigenmaps [98]. A brief description of these techniques is presented in the following paragraphs.

Kernel Principal Component Analysis (K-PCA)

Kernel PCA [95] is actually an extension of the conventional PCA in a high-dimensional space, which is obtained with the use of a kernel function. The main difference in comparison with the standard PCA is that the eigenproblem is solved for the "kernelized"

Table 6.2: Features extracted from each image in the database [6].

Cytoplasm Features	Nuclei Features
1. Area	1. Area
2. Brightness	2. Brightness
3. Short Diameter	3. Short Diameter
4. Longest Diameter	4. Longest Diameter
5. Elongation	5. Elongation
6. Roundness	6. Roundness
7. Perimeter	7. Perimeter
8. Maxima[1]	8. Maxima[1]
9. Minima[1]	9. Minima[1]
10. Nucleus Position	
11. Nucleus/Cytoplasm (size)	

[1] The number of pixels with the maximum/minimum intensity value in a 3 × 3 neighborhood of the specific area.

covariance matrix. If $X = \{x_1, x_2, ..., x_N\}$ is the original data set, the elements of this $N \times N$ matrix are defined as $k_{ij} = K(x_i, x_j)$, where K is the kernel function and x_i, x_j are D-dimensional feature vectors of X. In our implementation, we have used the polynomial and the Gaussian kernels. The kernel matrix is centered, in order the features in the high dimensional space to be defined by a kernel function with zero mean and the eigenvectors α_i are then calculated. The projection of a datum y_i in the low dimensional space is defined as:

$$y_i = \left\{ \sum_{j=1}^{N} a_1^j K(x_j, x_i), ..., \sum_{j=1}^{N} a_d^j K(x_j, x_i) \right\}, \tag{6.1}$$

where a_i^j denote the j-th component of the i-th vector and $d < D$ is the number of the retained eigenvectors.

Isomap

Isomap [96] is a variant of multidimensional scaling (MDS) [99], in which the distances between the datapoints in the high dimensional space are also retained in the low dimensional space. In MDS, this is accomplished by the eigendecomposition of a pairwise distance matrix (instead of the covariance matrix which is involved in PCA). In Isomap, the Euclidean distance between the points is substituted by their geodesic distance. Thus, the pairwise geodesic distance between the datapoints in the high dimensional space is preserved in the low dimensional space, by the construction of a neighborhood graph G, in which each datapoint is connected with its k nearest neighbors. The geodesic distance of two points may be approximated with the shortest path in the graph G between these points, using, for instance, Dijkstra's algorithm [100]. Having estimated the geodesic distances for all the points in the data set, the representation of the datapoints in the low dimensional space are computed by applying MDS on the resulting distance matrix.

Locally Linear Embedding (LLE)

This method is similar in spirit to Isomap, as it is also based on the construction of a distance graph G. However, in LLE [97] each datapoint is described as a linear combination of its k nearest neighbors, thereby assuming that the manifold is locally linear. The weights w_{ij} describe the contribution of the j-th point to the reconstruction of the i-th point and they are computed by minimizing the cost:

$$\arg\min_{W} E(W) = \sum_{i=1}^{N} ||x_i - \sum_{j=1}^{k} w_{ij} x_{i_j}||^2, \tag{6.2}$$

where x_{i_j} is the j-th nearest neighbor of the i-th point. Thus, the weights w_{ij} that best reconstruct each point x_i from its neighbors are used to compute the corresponding points y_i in the low dimensional space by minimizing the following cost function with respect to $Y = (y_1, y_2, ..., y_N)^T$:

$$\arg\min_{Y} \varphi(Y) = \sum_{i}^{N} ||y_i - \sum_{j=1}^{k} w_{ij} y_{i_j}||^2. \tag{6.3}$$

This minimization problem is equivalent to the calculation the eigenvectors corresponding to the smallest eigenvalues of the matrix $(I - W)^T(I - W)$, where I is the identity matrix and W is a matrix with elements w_{ij}. The above minimization is performed in two steps with the additional constraint $\sum_j w_{ij} = 1$ to make the representation translation invariant.

Laplacian Eigenmaps

The main philosophy of Laplacian Eigenmaps [98] is to calculate the low dimensional representation of the data in such a way that the local neighborhood information is optimally preserved. For this reason, the distance graph G is computed, in a way similar with the methods described above. Each edge of the graph is associated with a weight, which is a measure of closeness of the respective neighbors. The weights are attributed by the Gaussian kernel function $w_{ij} = e^{-\frac{||x_i - x_j||^2}{2\sigma^2}}$, where σ is the kernel width. Thus, the weights exhibit high values for nearest neighbors and small values for distant datapoints. Next, a diagonal matrix A is constructed, with elements $A_{ii} = \sum\limits_{j} w_{ij}, i = 1, .., n$ and the generalized eigendecomposition $Lu = \lambda Au$ is performed, where $L = A - W$. The low dimensional representation is obtained using the d eigenvectors corresponding to the smallest nonzero eigenvalues.

6.3 Results and discussion

In order to investigate the effectiveness of the above dimensionality reduction schemes, we have used two unsupervised classifiers and two datasets of patterns. More specifically, spectral clustering and fuzzy C-means are tested using patterns from two different feature sets (Table 6.1): one containing both cytoplasm and nucleus features (20 features) and the other containing only nucleus features (9 features). Several experiments were performed and the performance of the classification techniques was measured using patterns of increasing dimension varying from 1 to 20 features for the first subset and from 1 to 9 for the second subset. Furthermore, different values for the kernel width of spectral clustering have been tested $(10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 2, 5)$. In Isomap, LLE and Laplacian Eigenmaps, different numbers of nearest neighbors ranging from 4 to 20 were also tested for the construction of the distance graph G. The best results for each classifier are presented in this work.

For comparison purposes, PCA was also implemented. For the evaluation of the classification performance, the harmonic mean (HM) of the *Sensitivity* and the *Specificity* indices was calculated. The sensitivity measures the proportion of abnormal cells which are correctly identified as such by the classification algorithm, and the specificity measures the proportion of the normal cells that are correctly characterized as such.

The classification results of spectral clustering and fuzzy C-means are depicted in Table 6.3. For each feature subset, the HM and the number of features retained by the dimensionality reduction techniques are presented. As we can observe, the initial features without the use of dimensionality reduction schemes, lead to the weakest classification performance. The use of either linear or non-linear dimensionality reduction schemes results in a significant improvement of the classification.

More specifically, regarding the linear dimensionality reduction technique (PCA), we can conclude that there is a small improvement in the classification results, compared to the case where no dimensional reduction technique is used. Furthermore, in fuzzy C-means we observe a significant reduction in the retained features. Only 3 out of 20 dimensions are retained in first set of features and only 4 out of 9 for the nuclei feature subset. Finally, in spectral clustering, better classification results are produced when only the nuclei features are used, in comparison with the use of both nuclei and cytoplasm features.

In non-linear dimensionality reduction schemes, we can notice that the performance of the classifiers is clearly better when they are based only on nuclei features (except in the case of LLE with spectral clustering, where the results are approximately similar). In spectral clustering, an improvement of 12.16% in the classification is observed in Isomap, where the best value of HM (88.77%) using only the nuclei features is reached. Furthermore, in fuzzy C-means, the corresponding highest difference in classification rates is 11.17% and it is observed using K-PCA with polynomial kernel. Nevertheless, the best classification result using only the nuclei features is 90.58% with K-PCA and the Gaussian kernel. It must be noted that this result is obtained using only seven features, while for different number of features the HM value is smaller (Fig. 6.3).

The obtained results clarify that the use of non-linear dimensionality reduction schemes, not only improves the classification performance of spectral clustering and fuzzy C-means, but they also allow the successful separation of normal and abnormal cercival cells, based exclusively on nuclei features.

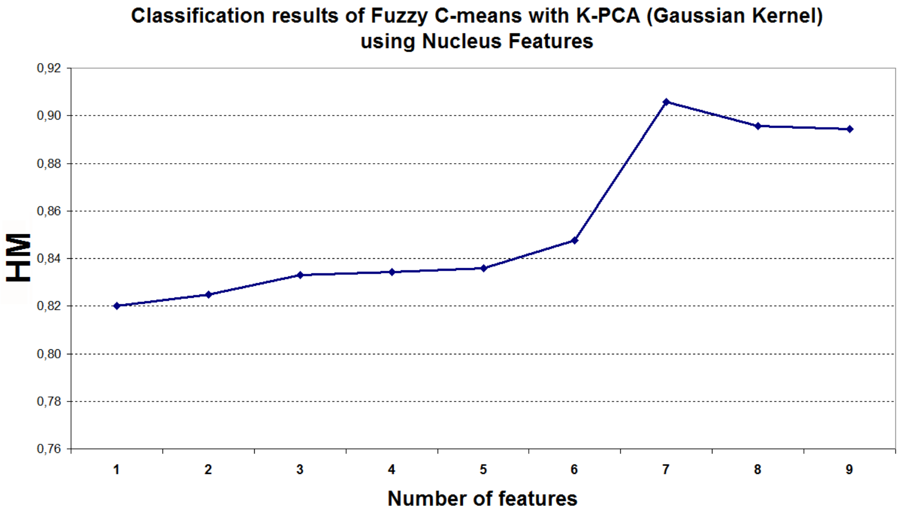


Figure 6.3: Results obtained in terms of HM for Fuzzy C-means classification. Notice that the HM reaches its highest value for seven features.

Table 6.3: Performance of classification in terms of HM and the number of retained features.

	Spectral Clustering				Fuzzy C-means			
	All Features		Nuclei Features		All Features		Nuclei Features	
	#feat	HM(%)	#feat	HM(%)	#feat	HM(%)	#feat	HM(%)
No dimensionality reduction	20	74.21	9	73.59	20	72.89	9	71.98
PCA	6	74.25	7	83.38	3	74.23	4	71.99
K-PCA (polynomial)	2	85.78	9	88.53	3	74.24	3	85.41
K-PCA (Gaussian)	16	84.44	7	87.52	9	**90.42**	7	**90.58**
Isomap	1	76.61	9	**88.77**	1	75.02	3	75.08
LLE	17	**86.97**	9	86.45	15	81.69	6	87.17
Laplacian Eigenmaps	20	80.84	3	87.52	11	85.31	1	87.20

6.4 Conclusion

The correct characterization of the cell nuclei in Pap smear images is a prerequisite for the derivation of accurate diagnostic decisions. Since in cell clusters presented in Pap smear images the automated cytoplasm segmentation is not feasible, in contrast to the automated nuclei segmentation [76, 77], we have investigated the case of the successful classification of cells with exclusively nuclei features using two unsupervised classifiers. In this direction, non-linear dimensionality reduction techniques were also used, for the more accurate representation of the features manifold. As it was verified by our experiments, the obtained results using only the nuclei features are better than the results obtained using all the extracted features (from the areas of nucleus and the cytoplasm). This implies that the characterization of a Pap smear image as normal or abnormal is feasible with the use of the nuclei features alone.

Chapter 7

Conclusions and Perspectives

In this thesis, we have studied in depth the special issues that Pap smear image processing exhibits. The accurate processing of these images that would lead to correct conclusions about the context of the examined slide is a difficult task for several reasons. The difficulties in the process of these images accrue from the noise and the artifacts that they contain, in combination with the extended cell overlapping and the variances in illumination and dye concentration of the cells. Thus, the effective process of such images demand specialized approaches that are able to provide object detection, object delineation, separation of partially occluded or overlapping objects and identification of normal and abnormal figures of the object of interest, which in the case of Pap smear images are the cells nuclei.

Our approach was mainly focused on the segmentation of the cells nuclei. This task is of high importance, as the nuclei detection and area determination result in the evaluation of salient nuclei features, which present great diagnostic value. Furthermore, we have extended the cell nuclei segmentation with a classification process for the detection of normal and abnormal nuclei, using exclusively shape and intensity nuclei features. More specifically, the techniques presented in this thesis concern the detection of the nuclei positions, the definition of the actual nuclei boundaries, the separation of the overlapped nuclei and finally the classification of normal and abnormal nuclei in Pap smear images. These techniques overcome the limitations arising from the complexity of these images, and they provide an effective framework for the reliable analysis of such images. This is clearly verified from the experimental results of the methods presented in this thesis. We can conclude that the proposed methodologies achieve accurate identification of the regions of interest and they present high performance, compared with the state of the art methods.

The method for the detection of the nuclei centroids described in Chapter 3 is based on mathematical morphology and clustering techniques and it includes *a priori* knowledge about the expected shape of the nuclei. This method results in the definition of the positions of the cell nuclei in a Pap smear image that contain both isolated cells and cell clusters. A rough estimation of the boundary of the nuclei was also calculated by the definition of the circumference of each nucleus, which resulted in the refinement of the detected centroid. This method can be directly applied to images captured from an optical microscope without the need of the predefinition of the region of interest. However, an issue that must be solved in the future is the separation of clustered nuclei, since the method in its current form indicates the existence of one nucleus in a specific location. Thus, an interesting extension of this method could be the annexation of an additional step, which would count the number of the overlapped nuclei in an eventual detected nuclei cluster.

In Chapter 4 we presented a method for the automated nuclei boundaries determination, which is able to extract the accurate boundaries of non overlapping nuclei. This method is accomplished with the application of the marker based watershed transform, using the nuclei markers extracted in the detection step. In order to include information from the three color channels of the image, the morphological color gradient image is calculated, in which the watershed transform is applied. The features of the detected areas concerning the shape, the texture and the image intensity are ranked through the mRMR feature selection scheme, in order to estimate their discriminative ability for the definition of the true nuclei set from the total findings. The features involving the neighborhood of the nuclei present high discriminative ability, indicating that useful information is contained not only inside the nucleus area but also in its neighborhood. The method was tested in a large image data set and it presents higher performance compared with segmentation methods based on deformable models (GVF, ACM). This method can be extended in the future with the investigation of the ability of the detected features for the recognition of normal and abnormal nuclei in Pap smear images.

Furthermore, a method for the boundary determination of overlapping nuclei is described in Chapter 5. The method incorporates *a priori* knowledge about the nuclei shape, and it is based on the training of a physically based deformable model in terms of modal analysis. Based on the estimated modal distribution and driven by the image characteristics the method succeeds in the determination of the nuclei boundaries in images containing two overlapping nuclei. The introduction of the locally adaptive image force using appropriate weight parameters, controls the contribution of the image force in the total energy of the deformable model. Thus, in the overlapping area, the deformation of the model is mainly driven by the learnt parameters obtained through the training procedure. Comparisons with other segmentation methods, proposed especially for the separation of overlapped nuclei, indicate that our method produces more accurate nuclei boundaries. In the future, the examination of the efficiency of our method in the segmentation of images containing three or more overlapped nuclei is an interesting research issue. Furthermore, the method can be extended with the development of an anterior step which would automatically determine the number of the overlapped nuclei.

In the above methods, the segmentation of cells nuclei in Pap smear images is efficiently addressed. The nucleus is the part of the cell which presents significant changes in abnormal circumstances and in contrast to the cytoplasm its segmentation is feasible in an automated manner. Thus, in Chapter 6, we investigated the case of the successful classification of cells in normal and abnormal categories based on features extracted exclusively from the nucleus area and ignoring the contingent cytoplasm features. We have examined non-linear dimensionality reduction schemes, in order to produce accurate representation of the features manifold and their influence on the classification performance of two unsupervised classifiers. The results indicate that high classification performance is achieved when only the nuclei features are used. In the future, experiments with supervised classifiers using exclusively nuclei features lying in low dimensional manifolds may provide an interesting extension of the proposed method.

The combination of all the methods described in this thesis could result in the development of an integrated fully automated system for the analysis of Pap smear slides, which would embody automated nuclei segmentation, nuclei feature extraction and finally classification. The methods described in this thesis are general and suitable for the addressing of the specific problems that Pap smear images processing presents. However, we can assert that their application is not limited only to these images, as the nuclei segmentation is the main field of research interest for the detection of the cancer in several cytological images. In the field of biomedical image analysis, the processing of microscopic images containing cells from several tissues, such as breast, lung or blood, occupies a large part of the scientific community. Thus, the proposed methods in this thesis, with the appropriate modifications, could be applied to a large range of microscopic images, resulting in the accurate segmentation and classification of the cells nuclei.

Appendix A

A. Fuzzy C-means clustering

The standard fuzzy C-means [62] objective function for partitioning a set of N unlabeled column vectors in R^p (where p is the number of features in each vector) $X_k, 1 \leq k \leq N$ into c clusters is given as:

$$J_m = \sum_{i=1}^{c} \sum_{k=1}^{N} \upsilon_{ik}^{m} \|x_k - v_i\|^2, \tag{A.1}$$

where $\{v_i\}_{i=1}^{c}$ are the prototypes of the clusters. The parameter m is the weighting exponent on each fuzzy membership, which determines the amount of fuzziness of the resulting classification. υ_{ik} represents the membership value of the feature vector x_k in cluster i. The following conditions must hold [62]:

$$\begin{gathered} 0 \leq \upsilon_{ik} \leq 1, \;\; i = 1,..,C, \;\; k = 1,..,N \\ \sum_{i=1}^{c} \upsilon_{ik} = 1, \;\; k = 1,..,N \\ 0 \leq \sum_{k=1}^{N} \upsilon_{ik} \leq N, \;\; i = 1,..,C \end{gathered} \tag{A.2}$$

The set of values satisfying the above conditions can be arranged in a matrix of the form $U[c, N]$. The fuzzy C-means objective function is minimized when high membership values are assigned to pixel data which are close to the centroid of its particular class. Low membership values are assigned to pixel data located far from the centroid. For the determination of crisp clusters, each column vector x_k is assigned to a cluster with maximum membership value.

B. Support Vector Machines

Given a training set $D = \{(x_i, y_i) | x_i \in R^p, y_i \in \{-1, 1\}\}_{i=1}^{n}$, where x_i are the data and y_i is the label for each x_i, the support vector machine (SVM) [63] determines the decision hyperplane between the two classes y_1 and y_2, which is obtained by the solution of the following optimization problem:

$$\min_{\boldsymbol{w},b,\xi} \left\{ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \xi_i \right\}$$
$$\text{subject to } y_i(\mathbf{w}^T\phi(x_i) + b) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \;\; i = 1..n \tag{A.3}$$

where $\mathbf{w}$ is a normal vector perpendicular to the hyperplane, C is a positive constant that reflects the influence of margin errors, b determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$ and ξ_i are slack variables, which measure the degree of misclassification of the datum x_i. The function $\phi(x_i)$ maps the training vectors (x_i) into a higher dimensional space. If we consider the dual form of this problem, the decision function is

$$sgn\left(\sum_{i=1}^{n}(y_i a_i K(x_i, x) + b\right), \tag{A.4}$$

where the terms $a_i, i = 1, .., n$ constitute a dual representation for the weight vector in terms of the training set, such as:

$$\mathbf{w} = \sum_i a_i y_i \mathbf{x}_i. \tag{A.5}$$

Furthermore, $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel function. In our experiments we have used the linear and the radial basis function (RBF) kernels, which are given by:

$$K(x_i, x_j) = x_i^T x_j, \tag{A.6}$$

and

$$K(x_i, x_j) = exp\left(-\gamma\|x_i - x_j\|^2\right), \;\; \gamma > 0, \tag{A.7}$$

respectively.

C. Spectral Clustering

Given a set of vectors $(x_1, x_2, .., x_N), x_k \in R^p$ and the number c of desired clusters to be separated, the spectral clustering algorithm [74] performs the following steps:

1. Define the affinity matrix $A^{N\times N}$ as $A_{i,j} = exp(-\|x_i - x_j\|^2/2\sigma^2)$.

2. Define the diagonal matrix $D^{N\times N}$ as $D_{ii} = \sum_i A_{ij}$

3. Define the matrix $L = D^{-1/2}AD^{-1/2}$.

4. Define the c largest eigenvalues $\lambda_i, i = 1, .., c$ of L and the corresponding eigenvectors $y_i, i = 1, .., c$.

5. Form the matrix Y which has as columns the eigenvectors y_i.

6. Normalize each row of Y to have unit length.

7. Treat each row of Y as a point in R^c and cluster them into c clusters via K-means.

8. Assign the original points x_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.

Appendix B

The watershed transform

The concept of watersheds [101] in image processing is based on considering an image in three dimensional space, with two spatial coordinates versus intensity. The value of the intensity is assumed to be the elevation information. In terms of this topographic representation of the image, the pixels are divided into three categories: pixels of regional minima, pixels of catchment basins and pixels of watershed lines, which separate neighboring catchment basins and consequently they separate different characteristic parts of the image. For the detection of the watershed lines in an image I with regional minima $M_1, M_2, ..., M_R$ a flooding process is performed in integer flood increments from $n_0 = \min(I) + 1$ to $n_{max} = \max(I) + 1$. Let $C(M_i), i = 1, .., R$ be the sets of points in the catchment basin corresponding to the regional minimum M_i and let $C[n]$ be the union of the flooded catchment basins at stage n of the flooding process. The set of the image points with intensity value lower than n is defined as $T[n] = \{p|I(p) < n\}$. The above sets of points are initialized as $C[\min(I) + 1] = T[\min(i) + 1]$. In the next steps of the algorithm, the set $C[n]$ is sequentially derived from $C[n-1]$ as follows:
Let Q be the set of the connected components in $T[n]$. Then for each connected component $q \in Q$ the intersection λ with the set $C[n-1]$ is calculated as $\lambda = q \cap C[n-1]$. Depending on the value of λ there are three possibilities:

1. If λ is empty then a new minimum is present and the connected component q is added into $C[n-1]$, thus $C[n] = C[n-1] \cup q$.

2. If λ contains one connected component of $C[n-1]$ then q belongs to an existing catchment basin of a regional minimum and consequently $C[n] = C[n-1] \cup q$.

3. If λ contains more than one connected component of $C[n-1]$ this means that q partially belongs to different catchment basins and the next step of flooding would cause the water level in these catchment basins to merge. For this reason, a watershed line must be constructed to prevent the overflow between these catchment basins.

The application of the watershed transform in this form usually results in oversegmentation of the image, because of the presence of artifacts and noise. To avoid this undesirable effect, the watersheds are applied in edge images with markers, which are connected components belonging to specific regions of interest in the image and they are used as starting points of the flooding process.

Appendix C

Local Binary Patterns

According to [3], the texture T in a local neighborhood of a monochrome image is the joint distribution of the gray levels of P ($P > 1$) image pixels:

$$T = t(g_c, g_0, .., g_{P-1}), \tag{C.1}$$

where the gray value g_c corresponds to the gray value of the center pixel of the neighborhood and $g_p(p = 0, .., P-1)$ correspond to the gray values of P equally spaced pixels on a loci of points (usually a circle with radius $R(R > 0)$). By subtracting the gray value of the center pixel g_c from the gray value of the neighborhood pixels, we obtain an equivalent form of the texture, that is:

$$T = t(g_c, g_0 - g_c, g_1 - g_c, .., g_{P-1} - g_c), \tag{C.2}$$

If we assume that the differences $g_p - g_c$ are independent of g_c, (C.1) can be factorized as:

$$T \approx t(g_c)t(g_0 - g_c, g_1 - g_c, .., g_{P-1} - g_c), \tag{C.3}$$

and since the distribution $t(g_c)$ describes the overall luminance of the image, it does not provide useful information for texture analysis, leading to a simplified form of (C.2)

$$T \approx t(g_0 - g_c, g_1 - g_c, .., g_{P-1} - g_c), \tag{C.4}$$

which is a highly discriminative texture operator, as it records the occurrences of various patterns in the neighborhood of each pixel in a P-dimensional histogram. The invariance with respect to the scaling of the gray scale is achieved by considering just the signs of the differences $g_p - g_c$ and not their exact value:

$$T \approx t(s(g_0 - g_c), s(g_1 - g_c), .., s(g_{P-1} - g_c)), \tag{C.5}$$

where

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases},$$

By assigning a binomial factor 2^p for each sign $s(g_p - g_c)$, (C.4) is transformed into a unique $LBP_{P,R}$ number that characterizes the spatial structure of the local image texture:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p. \tag{C.6}$$

It is observed that certain local binary patterns are fundamental properties of texture, and they are characterized as “uniform”. The uniformity measure $U(pattern)$ corresponds to the number of spatial transitions (bitwise 0/1 changes) in the pattern. In general, the operator for grayscale texture description using rotation invariant uniform patterns introduced by [3] is defined as:

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c), & \text{if } U(LBP_{P,R}) \leq 2 \\ P + 1, \text{otherwise} \end{cases} \tag{C.7}$$

where $U(LBP_{P,R}) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_P - g_c) - s(g_{p-1} - g_c)|$.

Bibliography

[1] O. Lezoray and H. Cardot, "Cooperation of color pixel classification schemes and color watershed: A study for microscopic images," *IEEE Transactions on Image Processing*, vol. 11, no. 7, pp. 783–789, 2002.

[2] L. Nanni, A. Lumini, and S. Brahnam, "Local binary patterns variants as texture descriptors for medical image analysis," *Artificial Intelligence in Medicine*, vol. 49, no. 2, pp. 117–125, 2010.

[3] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[4] C. Jung and C. Kim, "Segmenting clustered nuclei using H-minima transform-based marker extraction and contour parameterization," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2600–2604, 2010.

[5] C. Jung, C. Kim, S. Wan Chae, and S. Oh, "Unsupervised segmentation of overlapped nuclei using Bayesian classification," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 12, pp. 2825–2832, 2010.

[6] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear benchmark data for pattern classification," in *Proceedings of Nature inspired Smart Information Systems (NiSIS)*, 2005, pp. 1–9.

[7] G. N. Papanicolaou, "A new procedure for staining vaginal smears," *Science*, vol. 95, no. 2469, pp. 438–439, 1942.

[8] M. Arbyn, A. Anttila, J. Jordan, G. Ronco, U. Schenck, N. Segnan, H. Wiener, A. Herbert, and L. von Karsa, "European guidelines for quality assurance in cervical cancer screening," *Annals of Oncology*, vol. 21, no. 3, pp. 448–458, 2010.

[9] J. Monsonego, *Emerging numbers on HPV infections: from science to practice*, Karger, 2006.

[10] E. Artacho-Perula, R. Roldan-Villalobos, J. Salas-Molina, and R. Vaamonde-Lemos, "Histomorphometry of normal and abnormal cervical samples," *Analytical and Quantitative Cytology and Histology*, vol. 15, no. 4, pp. 290–297, 1993.

[11] M. L. Hutchinson, L. M. Isenstein, A. Goodman, A. A. Hurley, K. L. Douglass, K. K. Mui, F. W. Patten, and D. J. Zahniser, "Homogeneous sampling accounts for the increased diagnostic accuracy using the thinprep processor," *American Journal of Clinical Pathology*, vol. 101, no. 2, pp. 215–219, 1994.

[12] E. Davey, A. Barratt, L. Irwig, S. F. Chan, P. Macaskill, P. Mannes, and A. M. Seville, "Effect of study design and quality on unsatisfactory rates, cytology classifications, and accuracy in liquid based vs conventional cervical cytology," *Lancet*, vol. 367, no. 9505, pp. 122–132, 2006.

[13] M. Arbyn, C. Bergeron, P. Klinkhamer, P. Martin-Hirsch, A. G. Siebers, and J. Bulten, "Liquid compared with conventional cervical cytology," *Obstetrics & Gynecology*, vol. 111, no. 1, pp. 167–177, 2008.

[14] P. J. Klinkhamer, W. J. Meerding, P. F. Rosier, and A. G. Hanselaar, "Liquid based cervical cytology: A review of the literature with methods of evidence-based medicine," *Cancer Cytopathology*, vol. 99, no. 5, pp. 263–271, 2003.

[15] O. Abulafia, J. C. Pezzullo, and D. M. Sherer, "Performance of ThinPrep liquid-based cervical cytology in comparison with conventionally prepared Papanicolaou smears," *Gynecologic Oncology*, vol. 90, no. 1, pp. 137–144, 2003.

[16] R. L. Cahn, R. S. Poulsen, and G. Toussaint, "Segmentation of cervical cell images," *Journal of Histochemistry & Cytochemistry*, vol. 25, no. 7, pp. 681–688, 1977.

[17] H. Borst, W. Abmayr, and P. Gais, "A thresholding method for automatic cell image segmentation," *Journal of Histochemistry & Cytochemistry*, vol. 27, no. 1, pp. 180–187, 1979.

[18] E. Bengtsson, O. Erikssonand, and J. Holmquist, "High resolution segmentation of cervical cells," *Journal of Histochemistry & Cytochemistry*, vol. 27, no. 1, pp. 621–628, 1979.

[19] C. MacAulay and B. Palcic, "A comparison of some quick and simple threshold selection methods for stained cells," *Analytical and Quantitative Cytology and Histology*, vol. 10, no. 2, pp. 134–138, 1988.

[20] R. S. Poulsen and I. Pedron, "Region of interest finding in reduced resolution colour imagery-application to cancer cell detection," *Pattern Recognition*, vol. 28, no. 11, pp. 1645–1655, 1995.

[21] H. S. Wu, J. Gil, and J. Barba, "Optimal segmentation of cell images," in *Proceedings of the IEE Vision, Image and Signal Processing*, 1998, vol. 145, pp. 50–56.

[22] K. B. Kim, D. H. Song, and Y. W. Woo, "Nucleus segmentation and recognition of uterine cervical Pap-smears," in *Proceedings of the 11th RSFDGrC 2007, Lecture Notes in Computer Science*, 2007, vol. 4482, pp. 153–160.

[23] Z. Li and K. Najarian, "Biomedical image segmentation based on shape stability," in *Proceedings of the 14th IEEE International Conference on Image Processing (ICIP 2007)*, 2007, pp. 281–284.

[24] C. W. Chang, M. Y. Lin, H. J. Harn, Y. C. Harn, C. H. Chen, K. H. Tsai, and C. H. Hwang, "Automatic segmentation of abnormal cell nuclei from microscopic image analysis for cervical cancer screening," in *Proceedings of the 3rd IEEE International Conference on Nano-Molecular Medicine and Engineering*, 2009, pp. 77–80.

[25] M. H. Tsai, Y. K. Chan, Z. Z. Lin, S. F. Yang-Mao, and P. C. Huang, "Nucleus and cytoplast contour detector of cervical smear image," *Pattern Recognition Letters*, vol. 29, pp. 1441–1453, 2008.

[26] S. F. Yang-Mao, Y. K. Chan, and Y. P. Chu, "Edge enhancement nucleus and cytoplast contour detector of cervical smear images," *IEEE Transactions on Systems, Man and Cybernetics -Part B: Cybernetics*, vol. 38, no. 2, pp. 353–366, 2008.

[27] P. Malm and A. Brun, "Closing curves with Riemannian dilation: application to segmentation in automated cervical cancer screening," in *Proceedings of the 5th ISVC 2009, Lecture Notes in Computer Science*, 2009, vol. 5875, pp. 37–346.

[28] C. H. Lin, Y. K. Chan, and C. C. Chen, "Detection and segmentation of cervical cell cytoplasm and nucleus," *International Journal of Imaging, Systems and Technology*, vol. 19, no. 3, pp. 260–270, 2009.

[29] P. Bamford and B. Lovell, "A water immersion algorithm for cytological image segmentation," in *Proceedings of APRS Image segmentation workshop*, 1996, pp. 75–79.

[30] P. Jackway, "Morphological multiscale gradient watershed image analysis," in *Proceedings of the 9th Scadinavian Conference on Image Analysis (SCIA 1995)*, 1995, pp. 87–94.

[31] P. T. Jackway, "Gradient watersheds in morphological scale space," *IEEE Transactions on Image Processing*, vol. 5, pp. 913–921, 1996.

[32] K. B. Kim, S. Kim, and K. B. Sim, "Nucleus classification and recognition of uterine cervical Pap-smears using fuzzy art algorithm," in *Proceedings of 6th International Conference on Simulated Evolution and Learning, Lecture Notes in Computer Science*, 2006, vol. 4247, pp. 560–567.

[33] A. Kale and S. Aksoy, "Segmentation of cervical images," in *Proceedings of 20th International Conference on Pattern Recognition*, 2010, pp. 2399–2402.

[34] K. Nallaperumal and K. Krishnaveni, "Watershed segmentation of cervical images using multiscale morphological gradient and hsi colour space," *International Journal of Imaging Science and Engineering (IJISE)*, vol. 2, no. 2, pp. 212–216, 2008.

[35] N. Lassouaoui and L. Hamami, “Genetic algorithms and multifractal segmentation of cervical cell images,” in *Proceedings of 7th International Symposium on Signal Processing and its Applications*, 2003, number 2, pp. 1–4.

[36] E. Bak, K. Najarian, and J. P. Brockway, “Efficient segmentation framework of cell images in noise environments,” in *Proceedings of 26th Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2004, number 1, pp. 1802–1805.

[37] P. Sobrevilla, E. Montseny, and E. Lerma, “A fuzzy-based automated cell detection system for color Pap smear tests -FACSDS-,” *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, Studies in Fuzziness and Soft Computing*, vol. 220, pp. 657–674, 2008.

[38] F. Vaschetto, E. Montseny, P. Sobrevilla, and E. Lerma, “Threecond: An automated and unsupervised three colour fuzzy-based algorithm for detecting nuclei in cervical Pap smear images,” in *Proceedings of the 9th International Conference on Intelligent Systems Design and Applications*, 2009, pp. 1359 – 1364.

[39] N. Mustafa, N. A. Mat Isa, and M. Y. Mashor, “Automated multicells segmentation of thinprep ’ image using modified seed based region growing algorithm,” *Biomedical Soft Computing and Human Sciences*, vol. 14, no. 2, pp. 41–47, 2009.

[40] H. S. Wu, J. Barba, and J. Gil, “A parametric fitting algorithm for segmentation of cell images,” *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 3, pp. 400–407, 1998.

[41] A. Garrido and N. Perez de la Blanca, “Applying deformable templates for cell image segmentation,” *Pattern Recognition*, vol. 33, pp. 821–832, 2000.

[42] U. Grenander, *Pattern Synthesis, Lectures in Pattern Theory, Volume 1*, Springer-Verlag, New York, 1976.

[43] P. Bamford and B. Lovell, “Unsupervised cell nucleus segmentation with active contours,” *Signal Processing*, vol. 71, no. 2, pp. 203–213, 1998.

[44] N. M. Harandi, S. Sadri, N. A. Moghaddam, and R. Amirfattahi, “An automated method for segmentation of epithelial cervical cells in images of thinprep,” *Journal of Medical Systems*, vol. 34, pp. 1043–1058, 2010.

[45] E. Bengtsson, O. Eriksson, J. Holmquist, T. Jarkrans, B. Nordin, and B. Stenkvist, “Segmentation of cervical cells: Detection of overlapping cell nuclei,” *Computer Graphics and Image Processing*, vol. 16, pp. 382–394, 1981.

[46] D. Solomon, D. Davey, R. Kurman, A. Moriarty, D. O’Connor, M. Prey, S. Raab, M. Sherman, D. Wilbur, T. Wright, and N. Young, “The 2001 Bethesda System.

Terminology for reporting results of cervical cytology," *JAMA*, vol. 287, pp. 2114–2119, 2001.

[47] T. J. O'Leary, M. Tellado, S.-B. Buckner, I. S. Ali, A. Stevens, and C. W. Ollayos, "PAPNET-Assisted rescreening of cervical smears," *Journal of the American Medical Association*, vol. 279, no. 3, pp. 235– 237, 1998.

[48] S. F. Patten, J. S. J. Lee, and A. C. Nelson, "NeoPath, Inc. NeoPath AutoPap 300' Automatic Pap Screener System," *Acta Cytologica*, vol. 40, pp. 45–52, 1996.

[49] T. Chankong, N. Theera-Umpon, and S. Auephanwiriyakul, "Cervical cell classification using Fourier transform," in *Proceedings of 13th International Conference on Biomedical Engineering*, 2009, vol. 23, pp. 476–480.

[50] A. N. Bondarenko and A. V. Katsuk, "Extracting feature vectors of biomedical images," in *Proceedings of the 9th Russian-Korean International Symposium on Science and Technology (KORUS 2005)*, 2005, pp. 579 – 583.

[51] Y. Marinakis, G. Dounias, and J. Jantzen, "Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbour classification," *Computers in Biology and Medicine*, vol. 39, pp. 69–78, 2009.

[52] Y. Marinakis, M. Marinaki, and G. Dounias, "Particle swarm optimizaton for Pap-smear diagnosis," *Expert Systems with Applications*, vol. 35, pp. 1645–1656, 2008.

[53] L. H. Oliver, R. S. Poulsen, G. T. Toussaint, and C. Louis, "Classification of atypical cells in the automatic sytoscreening for cervical cancer," *Pattern Recognition*, vol. 11, pp. 205–212, 1979.

[54] R. F. Walker, P. Jackway, B. Lovell, and I. D. Longstaff, "Classification of cervical cell nuclei using morphological segmentation and textural feature extraction," in *Proceedings of the 2nd Australian and New Zeland Conference on Intelligent Information Systems*, 1994, pp. 297 – 301.

[55] D. Riana and A. Murni, "Performance evaluation of Pap smear cell image classification using quantitative and qualitative features based on multiple classifiers," in *Proceedings of the International Conference on Advanced Computer Science and Information Systems (ACSIS'09)*, 2009.

[56] N. A. Mat Isa, M. Y. Mashor, and N. H. Othman, "An automated cervical precancerous diagnostic system," *Artificial Intelligence in Medicine*, vol. 42, pp. 1–11, 2008.

[57] F. J. Gallegos-Funes, M. E. Gomez-Mayorga, J. L. Lopez-Bonilla, and R. Cruz-Santiago, "Rank M-type radial basis function (RMRBF) neural network for Pap smear microscopic image classification," *Apeiron*, vol. 16, no. 4, pp. 542–554, 2009.

[58] Z. Li and K. Najarian, "Automated classification of Pap smear tests using neural networks," in *Proceedings of the International Joint Conference on Neural Networks*, 2001, vol. 4, pp. 2899–2901.

[59] P. C. Huang, Y. K. Chan, P. C. Chan, Y. F. Chen, R. C. Chen, and Y. R. Huang, "Quantitative assessment of Pap smear cells by PC-based cytopathologic image analysis system and support vector machine," in *Proceedings of the 1st International Conference on Medical Biometrics*, 2007.

[60] J. Zhang and Y. Liu, "Cervical cancer detection using SVM based feature screening," in *Proceedings of 7th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'04)*, 2004, pp. 873–880.

[61] N. A. Mat, "Automated edge detection technique for Pap smear images using moving K-means clustering and modified seed based region growing algorithm," *International Journal of the Computer, the Internet and Management*, vol. 13, no. 3, pp. 45–59, 2005.

[62] J. C. Bezdek and S. K. Pal, *Fuzzy Models for Pattern Recognition*, New York, IEEE Press, 1992.

[63] N. Christianini and J. S. Taylor, *Support Vector Machines and other kernel-based methods*, Cambridge University Press, 2000.

[64] K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Image and Vision Computing*, pp. 474–485, 1994.

[65] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[66] L. Vincent, "Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms," *IEEE Transactions on Image Processing*, vol. 2, no. 2, pp. 176–201, 1993.

[67] E. J. Breen and R. Jones, "Attribute openings, thinings, and granulometries," *Computer Vision and Image Understanding*, vol. 64, no. 3, pp. 377–389, 1996.

[68] P. Soille, *Morphological Image Analysis: Principles and Applications*, Springer-Verlag, New York, 2003.

[69] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[70] C. Xu and J. Prince, "Snakes, shapes and gradient vector flow," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 359–369, 1998.

[71] K. H. Zhang, L. Zhang, H. H. Song, and W. Zhou, "Active contours with selective local or global segmentation: A new formulation and level set method," *Image and Vision Computing*, vol. 28, no. 4, pp. 668–676, 2010.

[72] A. N. Evans, "Morphological gradient operators for colour images," in *Proceedings of the IEEE International Conference on Image Processing (ICIP04)*, 2004, number 5, pp. 3089–3092.

[73] C. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[74] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, vol. 14, pp. 849–856, 2002.

[75] M. E. Plissiti, C. Nikou, and A. Charchanti, "Accurate localization of cell nuclei in Pap smear images using gradient vector flow deformable models," in *Proceedings of 3rd International Conference on Bio-inspired Signals and Systems (BIOSIGNALS)*, 2010, pp. 284–289.

[76] M. E. Plissiti, C. Nikou, and A. Charchanti, "Automated detection of cell nuclei in Pap smear images using morphological reconstruction and clustering," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 2, pp. 233–241, 2011.

[77] M. E. Plissiti, C. Nikou, and A. Charchanti, "Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images," *Pattern Recognition Letters*, vol. 32, no. 6, pp. 838–853, 2011.

[78] C. Zimmer and J. C. Olivo-Marin, "Coupled parametric active contours," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1838–1842, 2005.

[79] J. Cheng and J. C. Rajapakse, "Segmentation of clustered nuclei with shape markers and marking function," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 741–748, 2009.

[80] F. Cloppet and A. Boucher, "Segmentation of complex nucleus configurations in biological images," *Pattern Recognition Letters*, vol. 31, pp. 755–761, 2010.

[81] X. Bai, C. Sun, and F. Zhou, "Splitting touching cells based on concave points and ellipse fitting," *Pattern Recognition*, vol. 42, pp. 2434–2446, 2009.

[82] N. Malpica, C. Ortiz de Solorzano, J. J. Vaquero, A. Santos, I. Vallcorba, J. M. Garcia-Sagredo, and F. del Pozo, "Applying watershed algorithms to the segmentation of clustered nuclei," *Cytometry*, vol. 28, pp. 289–297, 1997.

[83] C. Nastar and N. Ayache, "Frequency - based nonrigid motion analysis: application to four dimensional medical images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 11, pp. 1067–1079, 1996.

[84] T. F. Cootes, C. J. Taylor, and J. Graham, "Active shape models - their training and application," *Computer Vision and Image Understanding*, vol. 1, no. 1, pp. 38–59, 1995.

[85] C. Nikou, G. Bueno, F. Heitz, and J. P. Armspach, "A joint physics-based statistical deformable model for multimodal brain image analysis," *IEEE Transactions on Medical Imaging*, vol. 20, no. 10, pp. 1026–1037, 2001.

[86] S. Krinidis and V. Chatzis, "A skeleton family generator via physics-based defoemable models," *IEEE Transactions on Image Processing*, vol. 18, no. 1, pp. 1 –11, 2009.

[87] S. Krinidis, C. Nikou, and I. Pitas, "Reconstruction of serially acquired slices using physics-based modeling," *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, no. 4, pp. 394–403, 2003.

[88] K. W. Wan, K. M. Lan, and K. C. Ng, "An accurate active shape model for facial feature extraction," *Pattern Recognition Letters*, vol. 26, no. 15, pp. 2409–2423, 2005.

[89] B. van Ginneken, A. F. Frangi, J. J. Staal, B. M. ter Haar Romeny, and M. A. Viergever, "Active shape model segmentation with optimal features," *IEEE Transactions on Medical Imaging*, vol. 21, no. 8, pp. 924–933, 2002.

[90] D. Shi, S. R. Gunn, and R. I. Damper, "Handwritten Chinese radical recognition using nonlinear active shape models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 277–280, 2003.

[91] A. Garrido and N. Perez de la Blanca, "Physically-based active shape models: initialization and optimization," *Pattern Recognition*, vol. 31, no. 8, pp. 1003–1017, 1998.

[92] G. Borgefors, "Hierarchical chamfer matching: a parametric edge matching algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 6, pp. 849–865, 1988.

[93] G. Borgefors, "Distance transformations in digital images," *Computer Vision, Graphics, and Image Processing*, vol. 34, no. 3, pp. 344–371, 1986.

[94] A. W Fitzgibbon, M. Pilu, and R. B. Fischer, "Direct least squares fitting of ellipses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 476–480, 1999.

[95] B. Scholkopf, A. Smola, and K. R. Muller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[96] J. C. Langford J. B. Tenenbaum, V. De Silva, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, pp. 2319–2323, 2000.

[97] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[98] M. Belikn and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[99] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, Chapman and Hall CRC, second edition, 2001.

[100] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, MIT Press and McGraw-Hill, second edition, 2001.

[101] R. C. Gonzalez and R. E. Woods, *Digital image processing*, Prentice Hall, second edition, 2002.

Author's Publications

Publications related to this thesis

Journal Publications

1. M. E. Plissiti and C. Nikou, "Overlapping cell nuclei segmentation using a spatially adaptive active physical model", submitted.

2. M. E. Plissiti, C. Nikou and A. Charchanti, "Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images", *Pattern Recognition Letters*, 32, 838-853, 2011.

3. M. E. Plissiti, C. Nikou and A. Charchanti, "Automated detection of cell nuclei in Pap smear images using morphological reconstruction and clustering", *IEEE Transactions on Information Technology in Biomedicine*, 15(2), 233-241, 2011.

Conference Publications

1. M. E. Plissiti and C. Nikou, "Cervical cell classification based exclusively on nucleus features", in *Proceedings of the International Conference on Image Analysis and Recognition* (ICIAR 2012), Aveiro, Portugal, 25-27 June 2012.

2. M. E. Plissiti and C. Nikou, "Cell nuclei segmentation by learning a physically based deformable model", in *Proceedings of the 17th International Conference on Digital Signal Processing* (DSP 2011), Corfu, Greece, 6-8 July 2011.

3. M. E. Plissiti, C. Nikou and A. Charchanti, "Watershed-Based segmentation of cell nuclei boundaries in Pap smear images", in *Proceedings of the IEEE International Conference on Information Technology and Applications in Biomedicine* (ITAB 2010), Corfu, Greece, 3-5 November 2010.

4. M. E. Plissiti, C. Nikou and A. Charchanti, "Accurate localization of cell nuclei in Pap smear images using Gradient Vector Flow deformable models", in *Proceedings of the 3rd International Conference on Bio-inspired Signals and Systems* (BIOSIGNALS 2010), 284-289, Valencia, Spain, 20-23 January 2010.

5. M. E. Plissiti, E. E. Tripoliti, A. Charchanti, O. Krikoni and D. I. Fotiadis, "Automated detection of cell nuclei in Pap stained smear images using fuzzy clustering",

in *Proceedings of the 4th European Congress for Medical and Biomedical engineering* (EMBEC 2008), 637-641, Antwerp, Belgium, 23-27 November 2008.

6. M. E. Plissiti, A. Charchanti, O. Krikoni and D. I. Fotiadis, "Automated segmentation of cell nuclei in Pap smear images", in *Proceedings of the IEEE International Conference on Information Technology and Applications in Biomedicine* (ITAB 2006), Ioannina, Greece, 26-28 October 2006.

Book Chapter

1. M. E. Plissiti and C. Nikou, "A review of automated techniques for cervical cell image analysis and classification", Biomedical imaging and computational modeling in biomechanics, Editors: U. Andreaus and D. Iacoviello, Springer, in press.

Other Publications

Journal Publications

1. C. V. Bourantas, I. C. Kourtis, M. E. Plissiti, D. I. Fotiadis, C. K. Katsouras and L. K. Michalis, "A method for 3D reconstruction of coronary arteries using biplane angiography and intravascular ultrasound images", *Computerized Medical Imaging and Graphics*, 29(8), 597-606, 2005.

2. C. V. Bourantas, M. E. Plissiti, D. I. Fotiadis, V. Protopappas, G. V. Mpozios, C. K. Katsouras, I. C. Kourtis and L. K. Michalis, "In vivo validation of a novel semi-automated method for border detection in intravascular ultrasound images", *British Journal of Radiology*, 78, 122-129, 2005.

3. M. E. Plissiti, D. I. Fotiadis, L. K. Michalis and G. Mpozios, "An automated method for lumen and media/adventitia border detection in a sequence of IVUS Frames", *IEEE Transactions on Information Technology in Biomedicine*, 8(2), 131-141, 2004.

Conference Publications

1. C. Bourantas, D. I. Fotiadis, I. C. Kourtis, L. K. Michalis and M. E. Plissiti, "Three-dimensional coronary artery reconstruction using fusion of intravascular ultrasound and biplane angiography", in *Proceedings of the International Congress and Exhibition of Computer Assisted Radiology and Surgery* (CARS 2003), London, 25-28 June 2003, International Congress Series, vol. 1256, 1133-1138, 2003.

2. C. Bourantas, D. I. Fotiadis, I. C. Kourtis, L. C. Kourtis, L. K. Michalis and M. E. Plissiti, "Quantitative validation of a 3D reconstruction automated method for coronary arteries", in *Proceedings of the 2nd European Medical and Biological Engineering Conference* (EMBEC 2002), Vienna, Austria, 4-8 December 2002.

3. C. Bourantas, M. E. Plissiti, D. I. Fotiadis and L. K. Michalis, “3D reconstruction of coronary arteries using intravascular ultrasound and biplane X-ray projections and its quantitative validation”, in *Proceedings of the XXIV Congress of the European Society of Cardiology* (ESC 2002), Berlin, Germany, 31 August- 4 September 2002.

4. C. Bourantas, M. E. Plissiti, D. I. Fotiadis and L. K. Michalis, “Validation of an automated method for the detection of the regions of interest in IVUS images”, in *Proceedings of the 23rd Panhellenic Cardiological Conference*, Athens, Greece, 2002.

5. M. E. Plissiti, D. I. Fotiadis and L. K. Michalis, “3D Reconstruction of arterial segments using IVUS and Angiographical Images”, in *Proceedings of the 6th Conference of Foundation for Research and Technology* (FORTH 2002), Ioannina-Metsovo, Greece, 1-3 March 2002.

6. M. E. Plissiti, D. I. Fotiadis and L. K. Michalis, “3D Reconstruction of stenotic coronary arterial segments using intravascular ultrasound and angiographic images”, in *Proceedings of the XVIIIth Congress of the International Society of Biomechanics* (ISB 2001), ETH Zurich, Switzerland, 8 - 13 July 2001.

7. C. Malizos, M. Hantes, A. Mavrodontides, M. Plissiti, D. I. Fotiadis, C. V. Massalas, A. Moukarika, T. Bakas and G. Evangelakis, “Quantitative monitoring and prognosis of osteogenesis”, in *Proceedings of the European Medical and Biological Engineering Conference* (EMBEC 1999), 37, 288-389, Vienna, Austria, 4-7 November 1999.

Book Chapter

1. A. Papadopoulos, M. E. Plissiti and D. I. Fotiadis, “Medical image processing and analysis for CAD systems”, in Medical Image Analysis Methods, CRC/Taylor and Francis Group, 2005.

Short Vita

Marina E. Plissiti was born in Ioannina, Greece in 1977. She received the B.Sc. and the M.Sc. degree from the Department of Computer Science, University of Ioannina, Greece, in 1998 and 2001 respectively. Since 2001 she is a secondary school teacher.

She has been dealt with several issues concerning the processing of biomedical images, such as X-RAY image processing (BSc Thesis), Intravascular Ultrasound (IVUS) image segmentation and 3-D reconstruction of coronary arteries through IVUS images and angiography (MSc Thesis), and microscopic image processing (PhD Thesis). Her research interests include medical image processing and artificial intelligence in biomedical applications.