

Supervised localization of cell nuclei on TMA images

Alessandro Ibba Robert P.W. Duin Marco Loog
Pattern Recognition Laboratory
Delft University of Technology
{a.ibba, r.p.w.duin, m.loog}@tudelft.nl

Abstract

We consider the problem of localizing renal cancer cell nuclei in Tissue Micro Array (TMA) images. We address this problem in three steps. An initial image processing-based procedure finds potential candidate nuclei, while the subsequent phase employs a trained classifier to prune the candidate cell nuclei found in the first.

A third phase is then used to perform a clustering of the positive classified blobs. In this work, we study cases when the second step is attained by extracting fixed size patches centred on the candidates, and representing these images with pixel-intensity histograms or related pair-wise distances (dissimilarities).

Our results, based on a Parzen classifier in the histogram feature space, show that the proposed procedure attains an optimal F1-measure of 0.9152 in localizing cell nuclei, providing state-of-the-art performance.

1. Introduction

Automatic localization of cell nuclei in tissue micro array (TMA) images is important for rapidly diagnosing cancer and at the same time to increase the accuracy of cancerous cell nuclei counts. The work presented proposes a procedure to automatically localize cell nuclei in TMA images. In our framework this goal is reached in three steps, where the second step makes use of pattern recognition techniques in order to improve the results achieved obtained at the first stage.

The reason for developing such a system is that one of the possible diagnosis methods for Renal Cell Carcinoma (RCC), one of the ten most frequent cancers in the EU [5] and the USA, is based on the histological tissue analysis, and currently the diagnostic rules are based on the precise counts of cancerous cell nuclei that are manually found and counted by pathologists. Our framework could be used to implement an almost fully automated cell nuclei localization in TMA images (fig. 1) where the experts would only have to confirm the results and provide a de-

cision only in special cases, instead of analysing each single TMA image. The proposed approach consists of three sub-phases, namely: a simple image-processing based technique employed to find potential candidates, a classification based phase aimed at pruning the found nuclei, and a post-processing phase aimed at a fusion of the positive candidates. Similar approaches are not new in similar medical imaging related and more specifically histo-pathological applications, employing not only approaches based on image-processing algorithms but also involving the use of statistical pattern recognition.

Procedures similar to the one used in this work have been employed for several different tissue types with images obtained with different markers and technology, some based on the Laplacian blob detector focused uniquely on the nuclei detection ([3]), others (like [10] [2]) have proposed a fully automated framework starting with the detection, then continuing with the segmentation and ending with a classification of the given images, while our focus lays on the localization of the cell nuclei present in the images in order to provide an additional instrument to the pathologists in the analysis of TMA images. An interesting contribution to the building of a benchmarking in the field of biomedical image analysis is proposed in [8].

With respect to the RCC images provided by the University of Zürich (and used in this paper) the studies of [7] and [6] have provided two different approaches in order to deal with the problem at hand but, although the overall scheme is to a certain extent comparable, the preprocessing and classification techniques chosen were totally different from the ones used in this work. Apart from the image-processing procedure used, in this paper we have also studied dissimilarity-based approaches that have been characterised by performances (in terms of F1-measure) significantly far from those obtained in this work (using Parzen), that as in [6] have reached a higher level than the one measured between the two pathologists (not equal to 1 due to the inter-pathologist variability).

2. Proposed Approach

2.1. Laplacian blob detector

Our preprocessing stage has been carried out using one of the most commonly used blob detectors based on the Laplacian of the Gaussian (LoG). Given an input image $I(x, y)$, this image is convolved by a Gaussian kernel

$$k(x, y, \sigma) = \frac{1}{2\pi\sigma} \exp\left(-\frac{(x^2 + y^2)}{2\sigma}\right)$$

at a certain scale σ in order to have the following scale-space representation:

$$L(x, y; \sigma) = k(x, y, \sigma) * I(x, y).$$

The Laplacian operator:

$$\nabla^2 L = L_{xx} + L_{yy}$$

is then computed, which usually leads to strong positive responses for low intensity (dark) blobs of extent and strong negative responses for high intensity (bright) blobs of similar size. A typical problem that is related to employing this type of operator at a particular scale is that its response is strongly dependent on the relationship between the sizes of the blobs present in the image and the width (σ) of the Gaussian kernel used in the pre-smoothing. In our specific problem we did not need to capture blobs of different (unknown) size in the image domain, since the variance of the cell nuclei sizes is quite limited, we therefore reached the conclusion that a combined (with respect to scale and space as in [13] and [14]) multi-scale approach is not necessary.

We have used images output of the Laplacian at different scales and thresholded at a percentage of the local maxima in order to retrieve the darker blobs (regions around the maxima) which should in principle correspond to the cell nuclei. The found blobs are then used in the following classification stage.

2.2. Feature extraction

The previously described procedure (2.1) has been used to find blobs that are then considered as potential cell-nuclei. For each of the found blobs centroids locations have been computed, and used to assign a label following a distance criterion, therefore the positive label has been assigned to those blobs whose centres were closer than a fixed threshold (average radius of the nuclei) d_r to any of the given polygons representing the real cell nuclei present in the tissue samples.

Using for instance all the image pixels to extract patches instead of using the found blobs would have lead to a much computationally heavier task and furthermore characterized by highly imbalanced classes. The described procedure (2.1) has been employed in order to reduce dramatically the

number of analysed patches while taking into account almost all the actual nuclei (found by the pathologists) therefore achieving close-to-one recall scores. The reason why we have chosen a distance based criteria and not for instance one based on the overlapping area is that this would have been a too strict constraint (leading to low levels of recall) while our initial goal was more to find many potential candidates or areas which may have contained a cell nuclei than obtaining a precise localization in the first stage.

The second phase aims at the pruning of the found candidates through means of a classification process. We have therefore extracted various features related to patches of fixed patch sizes (30, 40, 60 and 80 pixels) extracted around the centroids. These have been used in order to directly represent our objects (namely the found blobs) in the computed feature spaces or in related dissimilarity spaces. A third stage meant as a post-processing aims at further reducing the number of positively classified candidates through a clustering applied at the image level.

The features used in all our experiments have been the intensity histograms of the whole patches, considering different fixed binning sizes. While many publications ([12], [11] and [16]) suggest that an adaptive binning approach might be beneficial, our results do not generally show significant differences between bin sizes. In the presented experiments we have used 128 bins histograms.

2.3. Dissimilarity measures

Given two feature vectors \mathbf{x} and \mathbf{y} characterized by n components x_i and y_i which in our case have been intensity values, filters outputs or therefore related pixel-intensity histograms bins. On the basis of these feature spaces we have computed several distance measures that are briefly listed in this section:

- Euclidean distance:

$$d_{l_2}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- l_1 distance: $d_{l_1}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1 = \sum_{i=1}^n |x_i - y_i|$

- Earth Mover's Distance (EMD) in the closed form solution, applicable for instance in case of univariate histograms, where this measure is equivalent to the l_1 distance between histograms cumulative distributions.

$$d_{emd}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \left| \sum_{j=1}^i x_j - \sum_{j=1}^i y_j \right|$$

This measure has been effectively used in the work [16].

- Savitsky-Golay: this is a shape based distance measure between histograms and it is the l_1 distance between Gaussian derivatives of the spectra X (with a smoothing parameter σ)

$$X^\sigma = \frac{d}{di} G(i, \sigma) * X$$

where the histogram X can be seen as a feature vector \mathbf{x} and x_i its bins.

The shape-distance is defined as follows:

$$d_s(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n |x_i^s - y_i^s| \text{ (as further explained in [15]).}$$

2.4. Parzen classifier and its dissimilarity-based dual version

A simple linear classification algorithm in a dissimilarity spaces is highly non-linear in its original feature space. In particular a non-linear classifier like the Parzen windows can be approximated by a linear decision function applied in a ad-hoc constructed distance-space.

We consider in our experiments a nearest mean classifier in the dissimilarity space built using the distance measure derived from the Radial Basis Function (RBF) kernel:

$$k_{RBF}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{\sigma^2}\right)$$

Since an RBF kernel is semi-positive definite and therefore fulfils the Mercer's theorem we can derive a corresponding distance measure in the way that follows:

$$d_{RBF}^2 = k_{RBF}(\mathbf{x}, \mathbf{x}) + k_{RBF}(\mathbf{y}, \mathbf{y}) - 2k_{RBF}(\mathbf{x}, \mathbf{y})$$

that leads to:

$$d_{RBF}^2 = 2 - 2k_{RBF}(\mathbf{x}, \mathbf{y})$$

which is equivalent to:

$$d_{RBF}^2 = 1 - k_{RBF}(\mathbf{x}, \mathbf{y}) = 1 - \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{\sigma^2}\right) \quad (1)$$

Following the reasoning of ([17]), we consider a two class problem where N_+ objects are labelled positively ($y_i = +1$) and conversely N_- negatively ($y_i = -1$), computing the mean squared distances, in a RBF kernel space, with a mapping ϕ such that: $\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2^2 = -k(\mathbf{x}, \mathbf{x}')$ the label y_n of a test object \mathbf{x}_n would be defined as follows:

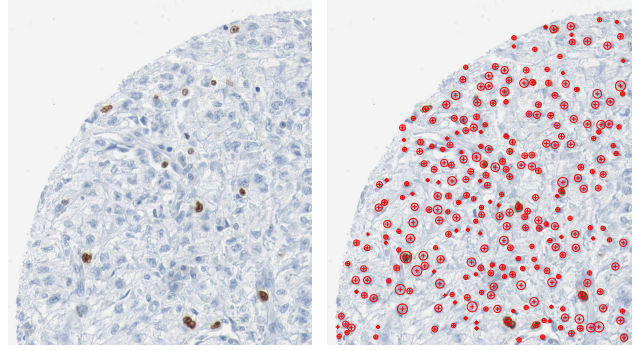


Figure 1. Left: Top left quadrant of a TMA spot from a ccRCC patient. Right: A trained pathologist localized and labelled the found cell nuclei. The size of each nuclei (according to the pathologist) correspond to the radius of the drawn polygon.

$$y_n = \text{sgn}\left(\frac{1}{N_+} \sum_{y_i=+1} k_{RBF}(\mathbf{x}_n, \mathbf{x}_i) - \frac{1}{N_-} \sum_{y_i=-1} k_{RBF}(\mathbf{x}_n, \mathbf{x}_i)\right) \quad (2)$$

The equation (2) can be seen as the Bayes decision on two Parzen windows estimates of the classes in the original feature space.

If we then substitute the kernel k_{RBF} using formula (1) with the RBF distance (d_{RBF}) we can eventually rewrite our classifier as follows:

$$y_n = \text{sgn}\left(\frac{1}{N_-} \sum_{y_i=-1} d^2(\mathbf{x}_n, \mathbf{x}_i) - \frac{1}{N_+} \sum_{y_i=+1} d^2(\mathbf{x}_n, \mathbf{x}_i)\right) \quad (3)$$

N.B.: this decision boundary is not equivalent to nearest mean classification algorithm in the dissimilarity space (defined in eq. (1)).

In this work we have tried to use other distance measure which might not show a euclidean behaviour. We have considered an RBF kernel where the squared l_2 distance has been substituted with the other three measures listed above in this section. From the so obtained kernels (which in some cases might not be semi-positive definite), the pairwise distances have been computed following the formula (1) in order to build different dissimilarity representations.

3. Experiments setup and performance measure

The conducted experiments can be described with the following steps. A Laplacian blob detector has been applied on the images using the following scales range $\sigma = [0.5 : 0.5 : 20]$ and therefore thresholded $th = [0.01 : 0.01 : 0.2]$ to binarize the filter outputs, and eventually extract the blobs. Patches of fixed sizes (30, 40 60 and 80 pixels), positioned on the blobs centroids, have been extracted from the original images, and labelled according to the distance (from the centroids indicated by the pathologist). On these patches features as described in section 1 have been extracted and related dissimilarity spaces have been computed, and employed to build several classifiers. The classification stage enables us to reduce significantly the number of potential cell nuclei, it is although not enough to be able to match them univocally to the given nuclei without losing in terms of precision. Therefore a clustering stage 3.1 (gaussian, hierachical) on the classifier output has been used in order to achieve a further pruning. The just mentioned post-processing phase (3.1) could in principle have taken place right after the extraction of blobs (first step), but by doing this the training set size would have ended up in being sensibly smaller, at the cost of the classification performance. Therefore we have chosen to apply this step only at the very end of the proposed framework.

Results of the experiments based on a leave one image out classification scheme (software used: [4]) are provided in terms of precision and recall per image (averaged over the 8 images) according to the one to one matching (meant as a linear assignment) performance measure described in section: 3.2.

3.1. Clustering

In order to reduce the number of blobs classified as nuclei we performed a clustering step on the image level. We therefore have been using four different approaches, the first three based on hierarchical clustering (single/complete/ average linkage) where the average nuclei radius d_r has been taken as distance threshold between the clusters centroids. The other methodology is based on a gaussian blurring (fig. 2.b) of the binary image of the (classified) blobs centra (fig. 2.a), which has been then thresholded (fig. 2.c) in order to obtain clustered centra. The standard deviation identical in both direction has been chosen as $\sigma = d_r/2$, we used the threshold $th = 0.99th^*$ where $th^* = 2/\pi d_r^2$ is the maximum value ($1/2\pi\sigma^2$) of a 2D gaussian with a covariance matrix $\Sigma = \sigma\mathbb{I}$. The 1D example shown in fig. 2.d explains the reasoning behind the choice of th . As soon as the mean of a gaussian curve (the last on the right side) is laying further than 2σ (d_r) from the closest curve, the chosen threshold enables us to distinguish it from the other curves, while

the three on the left side are clustered together as expected.

3.2. Performance measure

In order to evaluate the the detection performance we make use of the Precision-Recall space, therefore we need to count the true positives (TP), false positives (FP) and false negatives (FN).

A possible performance measure might take into account the view of a pathologist who in many cases might consider multiple matches candidate-nucleus allowing then for multiple true positives per actual cell nucleus. Therefore the false positive rate (FP) would correspond to the number of (by the classifier) positively labelled objects that at the same time are missing a match (according to the criterion) whereas each of the nuclei where the matched but negatively classified blobs together with the not matched cell nuclei would then be considered as false negatives. This choice allows for multiple matches because the matched blobs are indeed laying in the regions of real cell-nuclei thus considering them does not look like a sensible evaluation. The above explained measure, used for the candidates in the first stage, although simple and logic can not be easily compared with the results available in literature, we therefore chose an other measure based on a unique match between blobs and cell nuclei.

We have firstly applied a clustering stage to reduce the number of positively classified candidates (3.1) and then computed a distance matrix between the centra of the nuclei and those of the pruned blobs. The one-to-one match has then be determined employing a linear assignment algorithm (Hungarian method) [9], with respect to this step a major role is played by the chosen distance measure able to emphasize the distance of pairs outside the given nuclei radius (d_r) but at the same time keeping the within-radius euclidean distances unchanged. This distance d between the nucleus \mathbf{n}_i and the candidate \mathbf{b}_j has been defined as follows:

$$d(\mathbf{n}_i, \mathbf{b}_j) = \|\mathbf{n}_i - \mathbf{b}_j\|_2 \text{ if } d_{l_2}(\mathbf{n}_i, \mathbf{b}_j) < d_r \\ \text{else } d(\mathbf{n}_i, \mathbf{b}_j) = d^* \text{ where } d^* \gg \max_{i,j} d(\mathbf{n}_i, \mathbf{b}_j).$$

The linear assignment algorithm finds the cheapest (w.r.t. the overall distances sum) match between the two sets, the pairs with a distance smaller than the radius are eventually considered as positive matches (TP), the non matched candidates constitute the false positives (FP) and the non-found nuclei are counted as false negatives (FN).

We have then used the following performance measure based on precision ($P = TP/(TP + FP)$) and recall ($R = TP/(TP + FN)$), the F1-measure ($F = 2 PR/(P + R)$).

3.3. Data description

Our data consists of 8 separate RGB images of 1500 x 1500 pixels size with a corresponding labels file (fig: 1). Each image correspond to a TMA spot where the TMA-block was generated in a trial from the University Hospi-

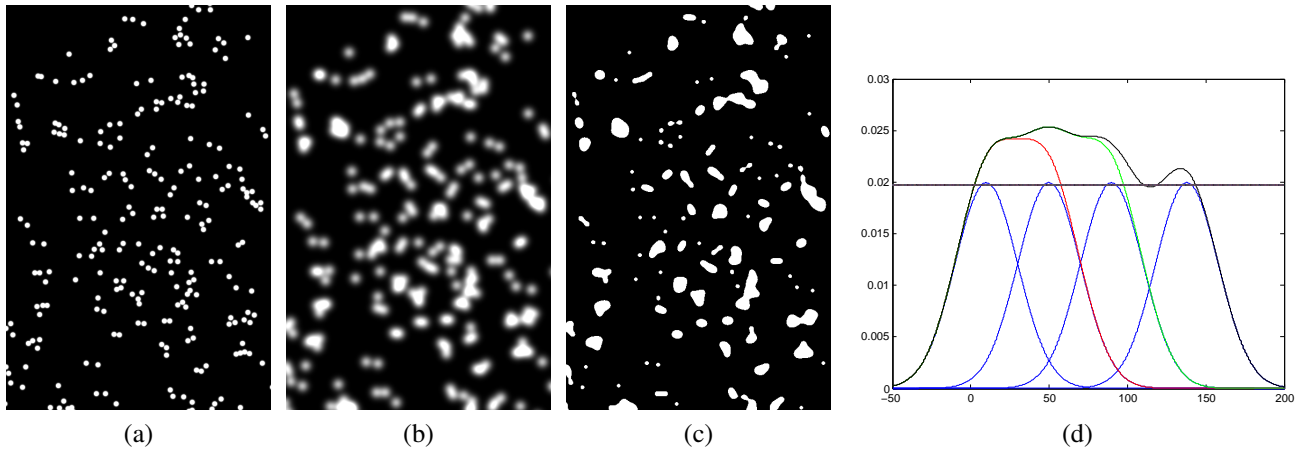


Figure 2. found blobs centroids (a), blurred image (b), thresholded image (c), 1D example (d)

tal Zürich. The TMA slides were immunohistochemically stained with the MIB-1 (Ki-67) antigen and scanned on a ScanScope virtual slide light microscope scanner from Aperio Technologies Inc. A lens with a magnification of 40x was used, which resulted in a per pixel resolution of 0.23 μm . Eventually the spots related to single patients were extracted as separate images of 1500 x 1500 x 3 pixels size. In order to produce the mentioned label files (which constitute our gold standard) two pathologists (from the University Hospital Zürich) experts in renal cell carcinoma used a software (developed at ETH Zürich) to annotate TMA spots of 8 different patients. They marked the location of each cell nucleus drawing a polygon around the nucleus centroid of its approximate radius size. In total each pathologist has detected more than 2000 cell nuclei on these images. For obvious reasons the experts are not providing exactly the same labelling (indicating the same cell nuclei with the same sizes), a measure of this (minimal) disagreement is therefore provided in section 4. For further details about the methodology and technological equipment used to retrieve the images used in this work we refer to [7].

We have eventually tested our procedure on Breast Cancer TMA images obtained from [1].

4. Results and discussion

The results in the form of Precision-Recall curves are given in figure: 5 show the two phases characterizing this study.

The first stage concerning the localization of the candidates with means of several blob detectors is represented by the cloud of blue '+' points where the scale (of the Laplacian 2.1) and the applied threshold have been varied in the way explained in 3. Therefore each single candidate has been labelled according to a distance (between the centroids of the found blobs and of the given nuclei) criterion. This

allows for counting the amount of true positives (TP), false positives (FP) and false negatives (FN) in order to compute precision and recall for that particular setting of the blob-detector. It is important to underline that the performance measure used at this point is based on the less strict one described in 3.2, while at this moment no classification stage takes place.

The choice of such a labelling criterion has been taken in order to allow for higher recall values, other criteria based for instance on the area overlap would have been much more restrictive therefore leading to smaller numbers of potential cell nuclei, but since our goal is to perform a pruning of the candidates this class of criteria has been avoided.

The blue '*' point represents the level of disagreement between the two pathologist considering one's nuclei as the ground truth and the other ones as candidates where the labelling criterion stayed unchanged. Those results concerning the first stage of our framework are constant in all the shown plots (fig. 5). Since our methodology heavily relies on the pruning of the determined blobs, we have chosen to focus the classification stage on those combinations of scale and thresholds which lead to high level of recall and at the same time not extremely low precision (red '+' points in the plots).

The red squares represents the classification results obtained training classifiers in feature spaces (or dissimilarity spaces) computed on patches (of fixed size) extracted around the centres of the found blobs. The green and black points represent respectively the best and the worst results in terms of F1-measure (harmonic mean of precision and recall), the underline (candidates) points are indicated in the same way. The procedure used to compute the used performance measure has been explained in the last part of section 3.2.

Of all the experiments described in 3 we reported the most relevant ones in fig. 5 where a Parzen classifier has

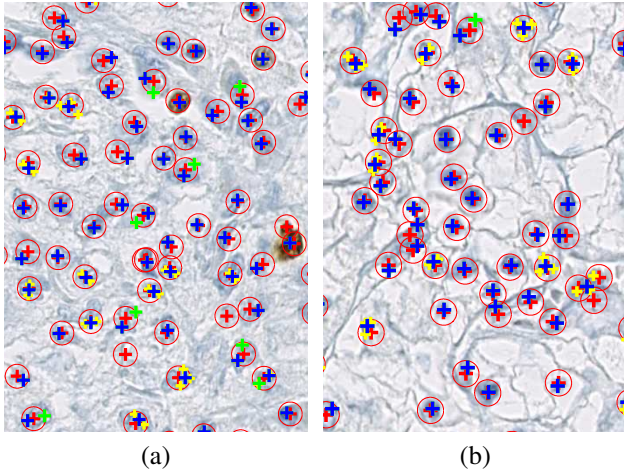


Figure 3. Two portions of two different RCC images where the blue '+' are the good classified centres the green are the false positive while the red circles are drawn around the nuclei.

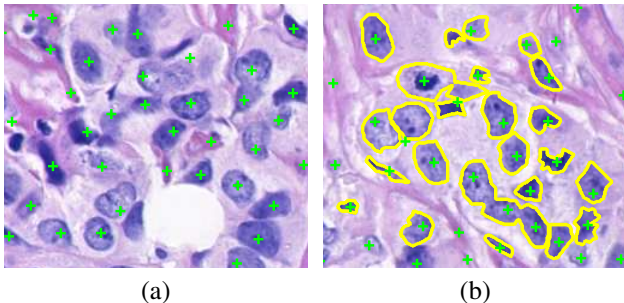


Figure 4. Two portions of a Breast cancer TMA image where the green '+' are the classifier outputs centres, in (b) it is possible to see the accuracy of the proposed methodology on the basis of the ground truth (yellow contours)

been trained in the original feature space. The proposed gaussian clustering (fig. 5.d) allows to reach a F1-measure equal to 0.9152. Those achieved utilizing the dissimilarity based approach have been characterised by significantly smaller values of F1. In fig. 3 two portions of two RCC TMA images are shown where the blue '+' correspond to the matched candidates (the yellow are candidates before the clustering stage), the green '+' are the unmatched results, while the red circles have been superimposed on the (by pathologists) given nuclei. We propose also an example where we applied the same methodology, object of this work, where this is applied on a different sort of TMA images (fig. 4) of which the ground truth was given. The image on the right side (fig. 4.b) shows the results in comparison with the superimposed ground truth, while the left image, where no evaluation is possible, is meant as an other example to be evaluated by experts.

5. Conclusions

Our results give an idea of the choices that we made in order to build and evaluate our classification frameworks. The disappointing results attained using the nearest mean classifier in the proposed dissimilarity spaces underline the higher non-linear characteristic of the problem at hand, which is moreover confirmed by the extremely good results achieved making of the Parzen windows classifier able to reach an F1-score higher than the one measured between the two pathologists. The results of the proposed framework on the RCC TMA images, are clearly better than those in [7] and comparable with those of [6] while the approach used in the latter appears to be far more complex than ours. We have also shown how well our method can perform on other types of TMA images.

References

- [1] UCSB biosegmentation benchmark, 2008. <http://bioimage.ucsb.edu/biosegmentation>.
- [2] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *Biomedical Engineering, IEEE Transactions on*, 57(4):841–852, 2010.
- [3] J. Byun, M. Verardo, B. Sumengen, G. Lewis, B. Manjunath, and S. Fisher. Automated tool for the detection of cell nuclei in digital microscopic images: application to retinal images. *Molecular vision*, 12:949–960, 2006.
- [4] R. Duin, P. Juszczak, P. Paclík, E. Pekalska, D. de Ridder, D. Tax, and S. Verzakov. Prtools (version 4.1.4). *Pattern Recognition Laboratory, Delft University of Technology, The Netherlands*, 2008.
- [5] J. Ferlay, P. Autier, M. Boniol, M. Heanue, M. Colombet, and P. Boyle. Estimates of the cancer incidence and mortality in europe in 2006. *Annals of oncology*, 18(3):581, 2007.
- [6] T. Fuchs and J. Buhmann. Inter-active learning of randomized tree ensembles for object detection. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1370–1377. IEEE, 2009.
- [7] T. Fuchs, T. Lange, P. Wild, H. Moch, and J. Buhmann. Weakly supervised cell nuclei detection and segmentation on tissue microarrays of renal clear cell carcinoma. *Pattern Recognition*, pages 173–182, 2008.
- [8] E. Gelasca, B. Obara, D. Fedorov, K. Kvilekval, and B. Manjunath. A biosegmentation benchmark for evaluation of bioimage analysis methods. *BMC Bioinformatics*, 10:368, 2009.
- [9] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.
- [10] K. Lee and W. Street. An adaptive resource-allocating network for automated detection, segmentation, and classification of breast cancer nuclei topic area: image processing and recognition. *Neural Networks, IEEE Transactions on*, 14(3):680–687, 2003.
- [11] W. Leow and R. Li. Adaptive binning and dissimilarity measure for image retrieval and classification. 2001.

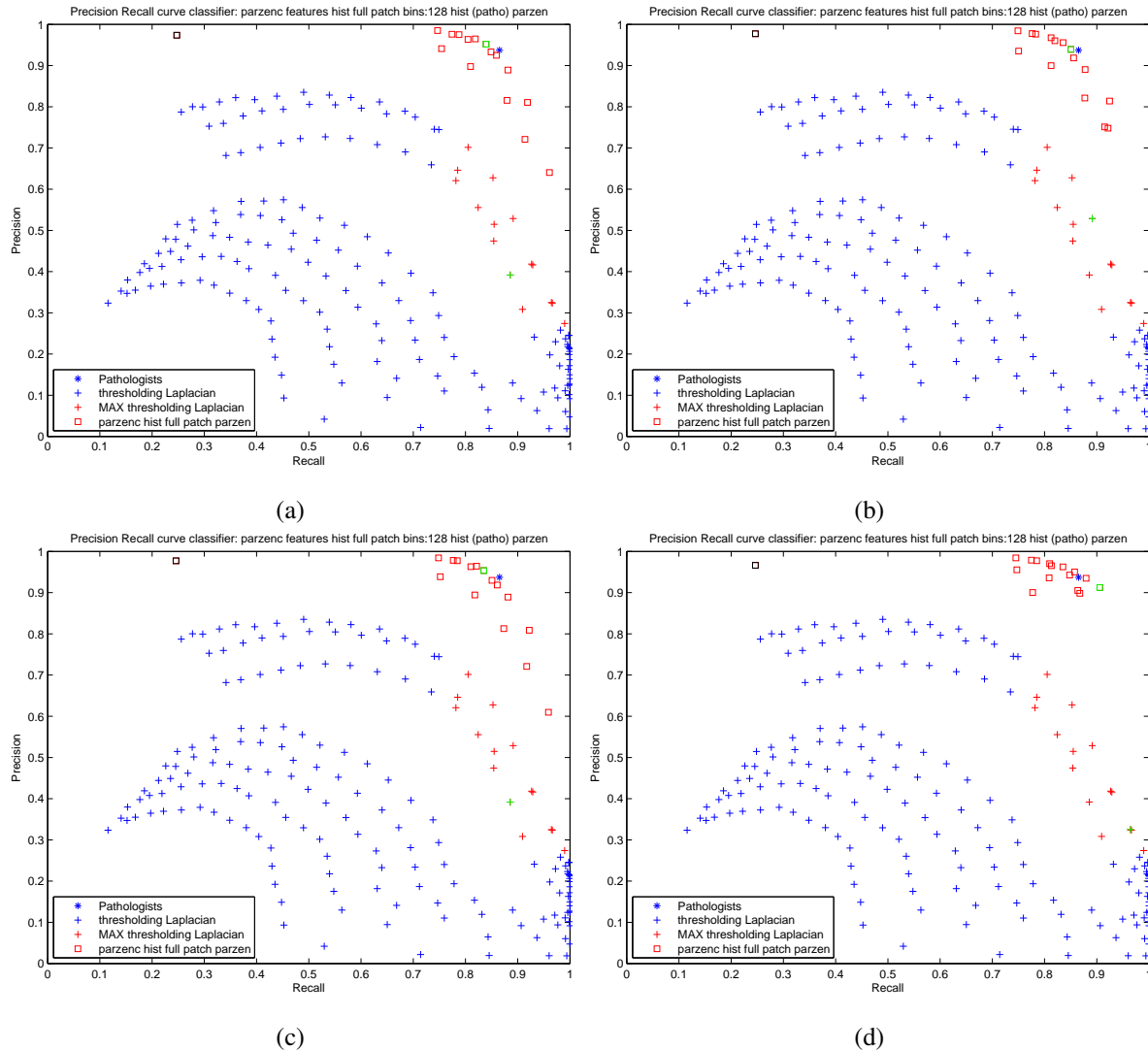


Figure 5. Precision-recall plots parzen classifier in the original feature space with hierarchical clustering applied on the output averaged (a) single (b) complete (c) and gaussian (d). Patch size: 30 pixel

[12] W. Leow and R. Li. The analysis and applications of adaptive-binning color histograms. *Computer Vision and Image Understanding*, 94(1-3):67–91, 2004.

[13] T. Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1):225–270, 1994.

[14] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[15] P. Paclík and R. Duin. Dissimilarity-based classification of spectra: computational issues. *Real-Time Imaging*, 9(4):237–244, 2003.

[16] J. Puzicha, J. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *Computer Vision, 1999. The Proceedings of the Seventh*

IEEE International Conference on, volume 2, pages 1165–1172. IEEE, 1999.

[17] B. Scholkopf. The kernel trick for distances. In *Advances in neural information processing systems 13: proceedings of the 2000 conference*, volume 13, page 301. The MIT Press, 2001.