Research Article

Training Nuclei Detection Algorithms with Simple Annotations

Henning Kost[1], André Homeyer[1], Jesper Molin[2,3,4], Claes Lundström[3,4], Horst Karl Hahn[1]

[1]Fraunhofer Institute for Medical Image Computing MEVIS, 28359 Bremen, Germany, [2]Department of Applied Information Technology, Chalmers University of Technology, 41258 Gothenburg, [3]Sectra AB, 58330 Linköping, Sweden, [4]Center for Medical Image Science and Visualization, Linköping University, 58183 Linköping, Sweden

Received: 10 January 2017 Accepted: 17 March 2017 Published: 15 May 2017

Abstract

Background: Generating good training datasets is essential for machine learning-based nuclei detection methods. However, creating exhaustive nuclei contour annotations, to derive optimal training data from, is often infeasible. **Methods:** We compared different approaches for training nuclei detection methods solely based on nucleus center markers. Such markers contain less accurate information, especially with regard to nuclear boundaries, but can be produced much easier and in greater quantities. The approaches use different automated sample extraction methods to derive image positions and class labels from nucleus center markers. In addition, the approaches use different automated sample selection methods to improve the detection quality of the classification algorithm and reduce the run time of the training process. We evaluated the approaches based on a previously published generic nuclei detection algorithm and a set of Ki-67-stained breast cancer images. **Results:** A Voronoi tessellation-based sample extraction method produced the best performing training sets. However, subsampling of the extracted training samples was crucial. Even simple class balancing improved the detection quality considerably. The incorporation of active learning led to a further increase in detection quality. **Conclusions:** With appropriate sample extraction and selection methods, nuclei detection algorithms trained on the basis of simple center marker annotations can produce comparable quality to algorithms trained on conventionally created training sets.

Keywords: Active learning, machine learning, nuclei detection, training set generation

Introduction

Many pathological assessments depend on the quantification of cell nuclei. In cancer diagnosis, for instance, the quantification of nuclei expressing the Ki-67 protein is a widely used method to determine the proliferation rate of a tumor. Furthermore, the quantification of lymphocytic infiltrates has been shown to be of strong prognostic importance.[1] Another important application is the determination of the progesterone and estrogen receptor status. The latter is arguably the most important predictive biomarker that exists today.[2] In clinical routine, such evaluations are usually done manually by estimating or counting a small number of nuclei, which is highly subjective and often not reproducible.[3] Consequently, the ability to automatically detect different types of nuclei on larger regions becomes increasingly important.

Varying staining and tissue preprocessing conditions, as well as different nuclear types and pathologies, lead to a huge variability in the appearance of nuclei, making their automatic detection very challenging. Recent approaches employ trainable algorithms to address this issue, including traditional machine learning[4-6] as well as deep learning methods.[7-9] Trainable detection methods come with the advantage of being adaptable and refinable by just using different training datasets.

Generating a good training dataset is essential for such methods. Most of these methods learn some kind of pixel-wise distinction between nuclear and nonnuclear regions,[4,6-9] to either create an intermediate segmentation or a probability map. Hence, the optimal training data would consist of exhaustive manual segmentations of all nuclei in several histological images. Unfortunately, creating such annotations requires an expert to accurately draw contour lines around each nucleus, making it a very tedious and time-consuming task.

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_3_17

Address for correspondence: Mr. Henning Kost, Fraunhofer Institute for Medical Image Computing MEVIS, Am Fallturm 1, 28359 Bremen, Germany. E-mail: henning.kost@mevis.fraunhofer.de

How to cite this article: Kost H, Homeyer A, Molin J, Lundström C, Hahn HK. Training nuclei detection algorithms with simple annotations. J Pathol Inform 2017;8:21.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2017/8/1/21/206227

Annotation marks at the nuclear centers constitute an alternative kind of reference data. Center annotations can be created with much less effort because they only require the expert to mark nuclei with a single click. This makes the marking process much faster and, therefore, also allows larger amounts of images to be annotated. Obviously, such annotations comprise much less information than full segmentations.

Center marker annotations have already been employed in the past. The different approaches address their insufficiency by augmenting them in various ways. An iterative thresholding approach was used by Gul-Mohammed *et al*.[10] to distinguish nuclear and nonnuclear areas around the center markers. In a study by Janowczyk and Madabhushi,[9] this distinction is performed by a naive Bayesian classifier with center positions as nuclear training data and randomly selected noncenter positions as nonnuclear training data. However, these approaches tie the capability of the machine-learning algorithm to the capability of the previous step. In both the studies by Sirinukunwattana *et al*.[11] and Xu *et al*.,[8] an assumption regarding the size of the nuclei is incorporated to supplement the annotation data: In a study by Sirinukunwattana *et al*.,[11] a regression is trained using the distance to the next center marker to compute the target value. In a study by Xu *et al*.,[8] nonnuclear training samples were drawn from positions that are further away from any center marker than a given threshold.

The quality of the mentioned approaches is hard to compare as the authors usually use different data sets with different nuclear types and often also different quality measures.

In a study by Vink *et al*.,[4] a nucleus detection method for Her2-stained breast tissue is proposed. The authors report a detection rate, which equals recall, of 0.95. Breast tissue nuclei are also detected in the studies conducted by Xing *et al*.[7] and Xu *et al*.[8] The approaches work on H&E-stained images and yield f1-measures of 0.78 and 0.84, respectively. In the study conducted by Arteta *et al*.[5] and Janowczyk and Madabhushi,[9] lymphocytic nuclei are detected in H&E-stained breast images. They state f1-measures of 0.88 and 0.90, respectively. Kårsnäs *et al*.[6] reported that a detection method for Ki-67-positive nuclei in breast tissue is proposed. The authors announce 1.0% missing objects, 2.6% missing annotations, and 4.1% multiple annotations. A nuclei detection method for H&E-stained colorectal tissue is described by Sirinukunwattana *et al*.[11] and an f1-measure of 0.80 is reported.

In this paper, we perform a systematic comparison of different methods for generating training sets solely from center marker annotations. In addition, we evaluate how the proposed center marker-based sample extraction methods compare with manual segmentations.

Methods

Training set generation

A training set consists of a set of training samples, which in turn consist of a feature vector and a class label. The content of the feature vector depends on the classification method that is to be trained. It might comprise hand-crafted features or in case of feature learning methods such as deep convolutional neural networks, small image patches. In both cases, a training sample is produced with respect to a given position in the image.

All of the examined training set generation approaches consist of two main steps, which are the extraction and the selection of training samples.

Given a set of training images with labeled center markers, the extraction step needs to identify image positions that can be labeled as nuclear or nonnuclear regions and derive a training sample from it. The main difficulty here is that center markers obviously provide far less information about the nuclear and nonnuclear regions in the image, especially with regard to their boundaries.

The output of the extraction step already forms a valid training set. However, the abundance of training samples often deteriorates the analysis quality and the runtime performance for the training process of the classifier. Depending on the type of the classifier, also, the runtime of the nuclei detection can be increased considerably. Thus, the second step of the considered training set generation approaches is the selection of optimal subsets of training samples.

Training sample extraction

We compare two different methods for extracting training samples from a given set of images.

Distance-based

We assume that positions close to the annotated center markers can be considered to represent nuclear regions whereas positions far away from any center marker are very likely to represent nonnuclear regions. Training samples are extracted as follows:

For each position x, y in an image, we compute the distance to the closest center marker. That distance and the index of that closest marker are stored in two maps d(x, y) and m(x, y). To be designated as nuclear region, a position must not be further away from the closest marker than a threshold called t_{nuc}. Thus, all positions x, y where $d(x, y) < t_{nuc}$ can be labeled as nuclear region. To be designated as nonnuclear region, a position may not be closer to any marker than a threshold called t_{bg}. Consequently, all positions x, y where $d(x, y) > t_{bg}$ are labeled as nonnuclear region. For our experiments, we set t_{bg} to 15 pixels and t_{nuc} to 3 pixels each at 20× resolution.

Voronoi-based

The distance-based approach has the drawback that the boundary positions of the nuclei are not considered at all. Boundary positions, however, are very informative because they shape the decision boundaries of the classifier. In our case, nuclear boundaries should be labeled as nonnuclear region so that clustered nuclei can be separated by the classifier. The Voronoi-based extraction method augments the distance-based method with such boundary samples.

The marker map m(x, y) is equivalent to the Voronoi diagram of the center markers. Assuming that neighboring nuclei are similarly sized, the Voronoi boundary between nontouching nuclei only crosses nonnuclear regions. As soon as two nuclei are touching, the Voronoi boundary crosses exactly that touching point. Consequently, for overlapping nuclei, the region of overlap is crossed by the Voronoi boundary. The assumption above may not always be valid, leading to Voronoi boundaries crossing nuclear regions, but we found that being a rare case in our experiments. Thus, the Voronoi boundaries are suited to extract nonnuclear samples along them.

Figure 1 illustrates the sample extraction methods.

Training sample selection

Selecting a subset of training samples from those extracted in the previous step can be beneficial. Reducing the amount of samples leads to a decrease of the runtime of the training process. For some classifiers, such as the random forest, the runtime of the classification is reduced as well.

Moreover, subsets of training samples often result in a higher quality of the nuclei detection if the samples show class imbalance. The extraction methods generally produce more samples of nonnuclear regions than nuclear regions because of the relative area fractions in the image. A small t_{nuc} further increases that imbalance. A classifier confronted with substantial class imbalance is deluged by instances of the majority class, leading it to ignore the instances of the minority class. Such imbalance is a well-known issue in the field of machine learning.[12]

During the training of a machine-learning classifier, the most interesting regions of the feature space are those close to the decision boundary of the classifier. Here, the classifier is most uncertain. That is why samples near the decision boundary are much more informative than samples far away. The ratio between samples of high and low informativeness in a training set can have a strong influence on the resulting detection quality. The samples extracted in the previous step are, in addition to the class imbalance stated above, likely to contain a large amount of uninformative instances.

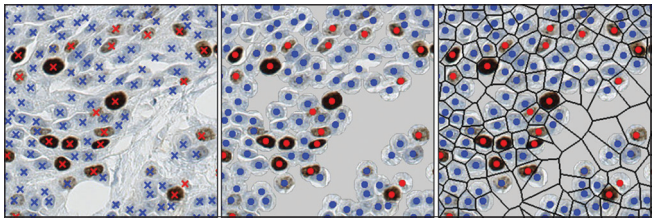

Figure 1: Visualization of the sample extraction methods. The left image shows the original image with overlayed center marker annotations. The center image shows the positions, where nonnuclear samples are extracted in gray and those where nuclear samples are extracted in red and blue for positively and negatively stained nuclei, respectively. The right image additionally shows the Voronoi boundaries in black, where also nonnuclear samples are extracted in the Voronoi-based extraction method

We investigated three different sample selection methods addressing the described issues.

Stratified random subsampling

The most straightforward way to reduce the amount of samples and to achieve a balance of the class labels is stratified random subsampling. From each class, samples are randomly drawn until a target number is reached or until there are no more samples of one class available. This method has the advantage that it can be integrated into the sample extraction methods. The subsampling can be already applied to the image positions before the features are calculated. This leads to a much better runtime performance than subsampling the samples in a separate step afterward.

Kd-tree subsampling

In the study by Pechenizkiy *et al*.,[13] the Kd-tree subsampling is suggested as an alternative or supplementary method for stratified random subsampling. It also reduces the number of samples while retaining their distribution in feature space. The general concept of the Kd-tree is explained by Bentley.[14] For the sample selection task, a Kd-tree with limited depth is constructed on the extracted samples using their features as dimensions. In each node, the splitting feature is chosen as that with maximum variance across the samples of the node and the median is used as pivot, as suggested by Omohundro.[15] Then, a single sample can be drawn randomly from each leaf of the tree. The granularity and the amount of resulting samples can be controlled by adjusting the depth limit of the tree. To also address class imbalance, we apply the Kd-tree subsampling independently for both classes and join the sample sets afterward.

Active learning

Active learning[16] selects samples with respect to their informativeness to the classifier. A classifier is trained using a subset S of the available samples. Then, iteratively, the remainder of the samples is classified and the classification confidence for each sample is considered. The samples with the least confident classifications are added to S for the next iteration. The iterations are terminated as soon as the size of S reaches a target number. By following this uncertainty sampling approach, the most informative samples are chosen from the training set. In our implementation, to produce a training set of n samples, the first subset is generated by randomly choosing $n/10$ samples from the available samples and $n/100$ samples are added in each iteration. In contrast to the previous methods, active learning does not address any class imbalance of the sample set.

The training sample selection methods are applied to either the samples extracted from a single image or the whole set of extracted samples. They can also be combined to utilize their different strengths.

Experimental setup

The different sample extraction and selection methods were compared using image data from a study described by Molin *et al*.[17] In that study, eight pathologists were asked to select

circular hot-spot regions containing approximately 200 nuclei from digitized Ki-67-stained breast tumor slides. From these hotspots, areas containing staining or scanning errors as well as overlapping areas were removed resulting in a set of 101 hot-spots from 24 different slides and cases. The digitized slides were downsampled if necessary to a magnification of 20×, and for each hotspot, a subimage containing that region was extracted. Center marker annotations for all nuclei within the circular hot-spot regions were created by a trained expert and verified by an experienced breast pathologist. Figure 2 examplary shows annotated hot-spot regions.

The evaluation is based on the nuclei detection method described by Kost *et al*.[18] A random forest assigns a probability value to each input image pixel for being close to the center of a nucleus. The feature set comprises:

- The normalized H, S, and V color channels
- The box filtered S channel
- An approximation of the difference of Gaussian on the S channel using box filters
- The radial symmetry on the S channel
- The box filtered radial symmetry.

Then, an optimized gray scale watershed algorithm is used to find and separate the individual nuclear regions. The algorithm is configured to only include positions with probability values above 0.5 into the nuclear regions as lower values indicate that it is more likely that the position belongs to background than to a nucleus. Another random forest uses the H, S, and V color channels to classify the staining within the nuclear regions and performs a majority vote to decide whether a nucleus is Ki-67 positive or negative.

To train the second classifier, we used modified versions of the sample extraction methods. For each position x, y with $d(x, y) < t_{nuc}$, an additional training sample for the second classifier was generated. The class label of the training sample was set depending on whether m (x, y) corresponds to a center marker of a Ki-67 positive or negative nucleus. This way, one training set was produced for each classifier. The selection methods were then applied to both sets individually using the same parameters.

The quality of the nuclei detection was assessed by comparing the results to the center marker annotations. Each detected nucleus was assigned to the closest center marker, provided that the distance of their positions was sufficiently close. A threshold of 10 pixels was found to be adequate. It corresponds to the approximate radius of the nuclei in the images. A one-to-one match was then considered a true positive (TP), a detected nucleus without a matching annotation was considered a false positive (FP), and an annotation without a matching detected nucleus considered a false negative (FN). In case when multiple detected nuclei were matched with the same center marker, one of these was considered TP whereas the others were counted as FP. Based on these values, precision, recall, and the f1-measure were computed as overall quality measures.

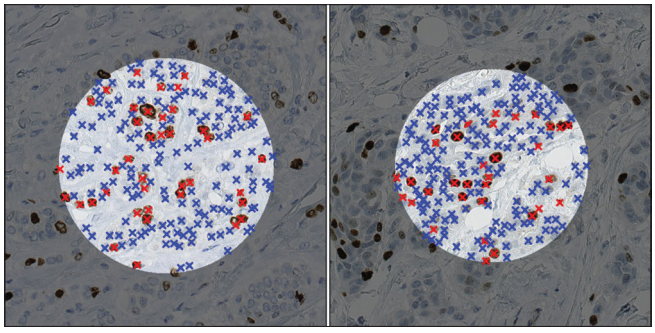

Figure 2: Visualization of the center marker annotations for two different images. The circle is scaled to contain approximately 200 nuclei. Inside the circle, all nuclei are annotated

For the experiments, we combined the sample extraction and selection methods in several ways to produce different training sets. The nuclei detection algorithm was then trained using these sample sets and the quality of the detection was assessed. To produce more robust results, the experiments were performed with 5-fold cross-validation. The folds were created in a way that images originating from the same slide were assigned to the same fold. This way, no training set is tested on the same slide it is created from. For all experiments, the same folds were used to ensure comparability. The quality measures of the individual folds were averaged to obtain the final measures.

Results

Experiment 1: Comparison of sample extraction and selection combinations

For the following experiment setups, distinct sample sets have been extracted by the distance-based and the Voronoi-based method. Then, different combinations of sample selection methods were applied to these sample sets. To obtain comparable results, all experiments, but (1a), which incorporates no selection method, produce a training set containing 2000 samples. This amount was found to produce adequate results while keeping the processing time for the classifier at an acceptable level. For the experiments involving two selection methods, the first one was applied per input image, leaving 256 samples per image. The latter one then was applied to the total of the remaining samples.

(a) No selection: As a base experiment, the outputs of the extraction methods were directly used to train the nuclei detection algorithm
(b) Random: The extracted samples were, as a whole, subjected to the stratified random subsampling selection method
(c) Kd-tree: The extracted samples were, as a whole, subjected to the Kd-tree subsampling selection method
(d) AL: The extracted samples were, as a whole, subjected to the active learning selection method
(e) Random + AL: The random subsampling selection method was applied per image. The final training set was then selected from the remaining samples using the active learning selection method
(f) Kd-tree + AL: The Kd-tree subsampling selection method was applied per image. The final training set was then

selected from the remaining samples using the active learning selection method

(g) AL + random: The effect of inverting the order of experiment (e) was examined. The active learning selection method was applied per image, followed by the stratified random subsampling selection method

(h) AL + Kd-tree: The effect of inverting the order of experiment (f) was examined. The active learning selection method was applied per image, followed by the Kd-tree subsampling selection method

(i) Kd-tree + random: In this experiment, both class balancing methods were combined. The Kd-tree subsampling selection method was first applied per image, and then the stratified random subsampling selection method was applied on the remaining samples afterward

(j) AL + AL: In this experiment, the active learning selection method was applied to both the per-image and the remaining samples.

Table 1 shows the results of the described experiment setups. First of all, we can state that the Voronoi-based extraction method yields quality measures slightly superior to the distance-based method in most experiment setups. Looking at the selection methods, we can see that the experiments that do not comprise a class balancing method lead to far worse quality measures. This can be observed in experiments (1d) and (1j), which only consist of active learning, and especially in experiment (1a), where no selection is performed at all. The best results are obtained by the combinations that include class balancing and active learning.

The tested sample extraction methods produce highly imbalanced training sets. On average, only 6.09% or 5.54% of the samples belong to the nuclear class for the distance-based and Voronoi-based extraction, respectively. The imbalance affects the resulting classification in a negative way. This can be observed in experiment (1a). The detection quality for the unprocessed training sets is low.

The usage of active learning alone, as shown in experiments (1d) and (1j), does improve the detection quality slightly but still yields results well inferior to other experiments. This indicates that active learning is not very well suited to deal with these large imbalances, which stems from the way active learning selects new samples. When there are mostly nonnuclear samples to choose from, the most uncertain samples are likely to be imbalanced toward nonnuclear samples as well. For this reason, a proper balancing of the samples is advisable.

The absence of class balancing in experiments (1a), (1d), and (1j) results in a strong bias of the classifier, which can be observed as a considerable difference in the precision and recall values. In experiment 2, precision-recall-curves (PR-curves) are analyzed to further examine this issue.

The stratified random subsampling and the Kd-tree-based selection seem to be equally suited for balancing as

Table 1: Quality measures of both proposed sample extraction methods combined with different sample selection methods

	Ki-67-positive nuclei			Ki-67-negative nuclei			All nuclei					
	TP	FP	FN	TP	FP	FN	TP	FP	FN	Precision	Recall	f1-measure
Distance-based												
(a) No selection	650.8	74.0	189.8	803.6	239.2	2024.0	1454.4	313.2	2213.8	0.823	0.396	0.530
(b) Random	716.8	156.8	123.8	2220.6	500.6	607.0	2937.4	657.4	730.8	0.817	0.801	0.806
(c) Kd-tree	721.6	174.0	119.0	2203.0	509.4	624.6	2924.6	683.4	743.6	0.811	0.797	0.801
(d) AL	688.6	66.2	152.0	1658.0	232.0	1169.6	2346.6	298.2	1321.6	0.887	0.640	0.740
(e) Random + AL	732.8	161.2	107.8	2290.4	556.2	537.2	3023.2	717.4	645.0	0.808	0.824	0.814
(f) Kd-tree + AL	732.6	151.2	108.0	2246.6	535.0	581.0	2979.2	686.2	689.0	0.813	0.812	0.810
(g) AL + random	706.2	118.6	134.4	2176.8	451.4	650.8	2883.0	570.0	785.2	0.835	0.786	0.807
(h) AL + Kd-tree	715.8	107.4	124.8	2202.8	460.4	624.8	2918.6	567.8	749.6	0.837	0.796	0.813
(i) Kd-tree + random	718.2	192.2	122.4	2188.6	516.6	639.0	2906.8	708.8	761.4	0.804	0.792	0.795
(j) AL + AL	683.4	76.6	157.2	1964.2	369.4	863.4	2647.6	446.0	1020.6	0.856	0.722	0.781
Voronoi-based												
(a) No selection	611.8	66.6	228.8	585.0	151.4	2242.6	1196.8	218.0	2471.4	0.846	0.326	0.467
(b) Random	746.0	206.4	94.6	2333.6	558.4	494.0	3079.6	764.8	588.6	0.801	0.840	0.817
(c) Kd-tree	750.4	192.6	90.2	2294.2	506.4	533.4	3044.6	699.0	623.6	0.813	0.830	0.819
(d) AL	628.8	50.8	211.8	1535.0	208.0	1292.6	2163.8	258.8	1504.4	0.893	0.590	0.711
(e) Random + AL	761.8	192.6	78.8	2364.8	561.8	462.8	3126.6	754.4	541.6	0.806	0.852	0.826
(f) Kd-tree + AL	763.2	188.0	77.4	2364.2	576.0	463.4	3127.4	764.0	540.8	0.804	0.853	0.825
(g) AL + random	728.8	119.4	111.8	2210.6	458.6	617.0	2939.4	578.0	728.8	0.836	0.801	0.815
(h) AL + Kd-tree	726.8	120.4	113.8	2195.8	444.0	631.8	2922.6	564.4	745.6	0.838	0.797	0.814
(i) Kd-tree + random	744.6	176.6	96.0	2314.8	541.4	512.8	3059.4	718.0	608.8	0.810	0.834	0.819
(j) AL + AL	694.8	91.4	145.8	1734.4	288.2	1093.2	2429.2	379.6	1239.0	0.865	0.662	0.747

TP: True positive, FP: False positive, FN: False negative, AL: Active learning

the comparison of the quality measures in experiments (1b) and (1c), (1e) and (1f), as well as (1g) and (1h) indicates. However, since stratified random subsampling is much simpler and improves the runtime performance when integrated into the extraction step, it is to be preferred over the Kd-tree-based approach. The best results were achieved by experiment setup (1e) being the combination of Voronoi-based extraction, stratified random subsampling, and active learning. Two example outputs are visualized in Figure 3. Another interesting approach is (1b), the solely applied stratified random subsampling. It is simple, yields good results, and has a good runtime performance due to the integrability into the sample extraction step. However, in general, the differences of the methods that use balancing are rather small. In contrast, the differences between the methods with and without balancing are major.

Experiment 2: Precision-recall-curves

As described in section 2.2, a cutoff value of 0.5 was used for the experiments, which is the natural threshold for a two-class problem. However, it is interesting to investigate how different cutoff values influence precision and recall.

For each approach described in experiment 1, the cutoff value was altered in 16 steps between 0 and 1, and at each step, precision and recall were determined. Figure 4 shows the PR-curves for all approaches in an overview graph. In the subsequent graphs, the curves are reduced and grouped to highlight different aspects. Furthermore, the axes are scaled to only show the most interesting quadrant of the graph.

In Figure 5, the PR-curves are divided into approaches that contain a sample selection method and approaches that do not, which is only the case for (1a) curves. It is clearly visible that the application of even the most basic sample selection methods improves the quality of the nuclei detection considerably. This is the case for both sample extraction methods.

Figure 6 shows the approaches that contain sample selection grouped according to their sample extraction methods. Here, it becomes apparent that the Voronoi-based extraction leads to better results than the distance-based extraction. This is especially the case for recall.

In Figure 7, only the approaches using the Voronoi-based sample extraction are plotted. We found that approaches consisting of two subsequent sample selection methods with at least one of them incorporating active learning perform especially well. These approaches are highlighted in this figure. Active learning per image followed by a selection that performs class balancing (1g) and (1h) leads to the best results for both extraction methods.

The PR-curves shown in this section have an unusual shape. Normally, with cutoff values becoming lower, the precision declines while the recall grows toward 1. In our case, the recall does not increase after a certain value but decreases again. The reason for this behavior is the watershed algorithm which is part of the nuclei detection method. This limits the number of detected nuclei. With a low cutoff value, more pixel positions

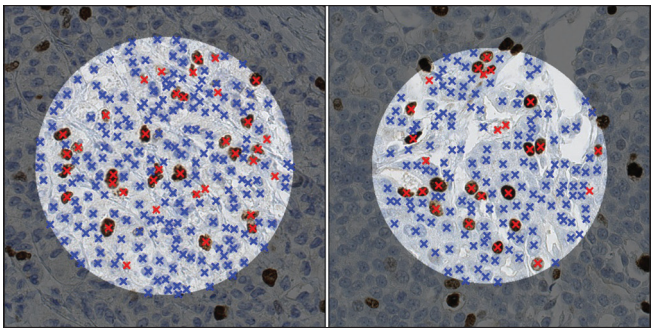

Figure 3: Example results of experiment (1e). The red and blue markers show Ki-67 positive and negative nuclei as detected by the algorithm, respectively

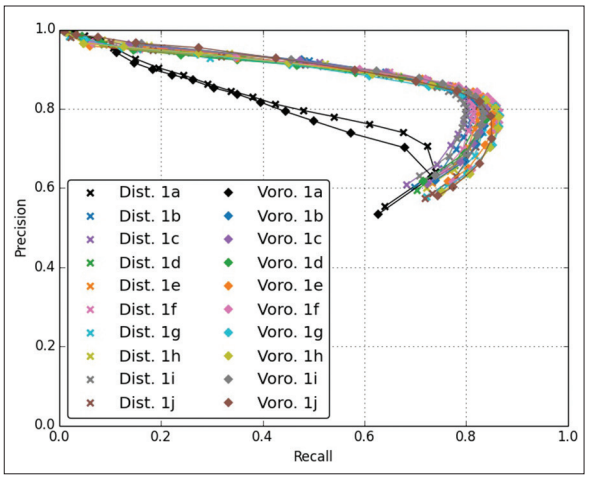

Figure 4: Overview plot of the precision-recall-curves for all training approaches and both distance-based (dist.) and Voronoi-based (voro.) sample extraction. The labels 1a–1j correspond to the notation in experiment 1

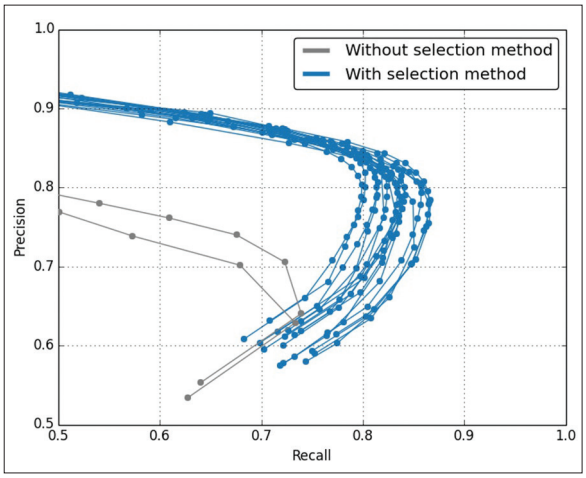

Figure 5: Precision-recall-curves showing the quality improvements when using a sample selection method (blue) compared to approaches without sample extraction (gray)

are being considered by the algorithm. Nevertheless, those are likely to be assigned to an existing nuclear region instead of constituting a new region. Another effect is that the nuclear

regions, which are segmented by the watershed algorithm, become larger. The nuclear positions are computed as the center points of the nuclear regions and the positivity of the nuclei is derived from the staining classification results within the nuclear regions. When such regions become unreasonably large, nuclear positions or their positivity might become incorrect. This effect causes the decrease of recall at low cutoff values.

Experiment 3: Impact of the training set size

The impact of the training set size on the quality of the nuclei detection was evaluated in experiment 3. Training sets of different sizes were produced using the Voronoi-based extraction method, followed by a selection method as described in (1b) and (1e), which appeared to be the most interesting approaches in experiment 1. For the latter approach, the active learning was parametrized to select 10% of the samples selected by the stratified random subsampling, which is comparable to the ratio in the above experiments. The overall f1-measure of the training sets was then assessed to compare the learning curves of these two methods. Training set sizes from 100 up to 5000 samples have been evaluated with an offset of 100 and up to 15,000 samples with an offset of 1000.

Figure 8 shows the results of experiment 3. Both learning curves have an approximately asymptotic shape. They rise steeply until about 2000 samples and ascend more slowly afterward. Nevertheless, the learning curve for the approach containing active learning shows superior quality values throughout all training set sizes. The experiment shows that the number of 2000 samples for a training set is a reasonable choice. Although more samples would slightly increase the quality of the nuclei detection, we consider this a good compromise between quality and runtime performance.

Experiment 4: Comparison of extraction methods with manual segmentations

To assess the quality of the sample extraction methods, we compared a manual nuclei segmentation with the proposed distance and Voronoi-based extraction methods. To produce a training set from segmentation annotations, nonnuclear samples were generated from all positions outside nuclear regions. Since the trained method should yield maximum nucleus probability at the center of the nuclei, the nuclear samples were only generated at the centers of the nuclei, equally to the extraction methods proposed.

As the selection method, we used stratified random subsampling per image followed by active learning (1e), which appeared to achieve the best results in experiment 1. We used ten images with exhaustive nuclei segmentation annotations which do not belong to the image set used in experiment 1–3. Samples were extracted from the annotations for each pixel. To compare those samples with the described extraction methods, the segmentation annotations were reduced to center markers by computing the center of gravity of each segment. For this experiment, we did not perform a cross-validation but tested the resulting training sets using the image set described above. The

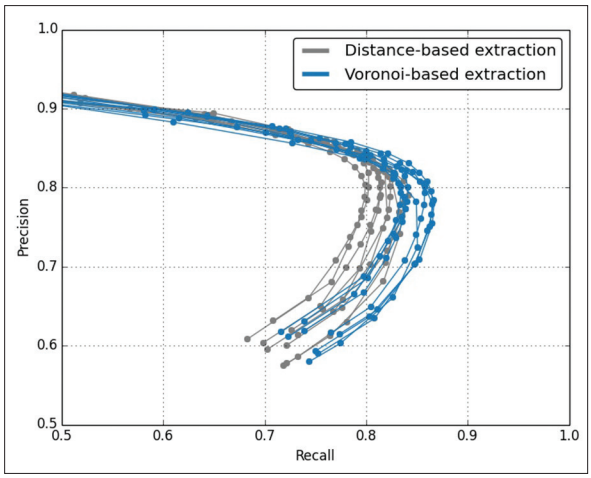

Figure 6: In most cases, the approaches using Voronoi-based sample extraction (blue) lead to better detection quality than those using distance-based sample extraction (gray)

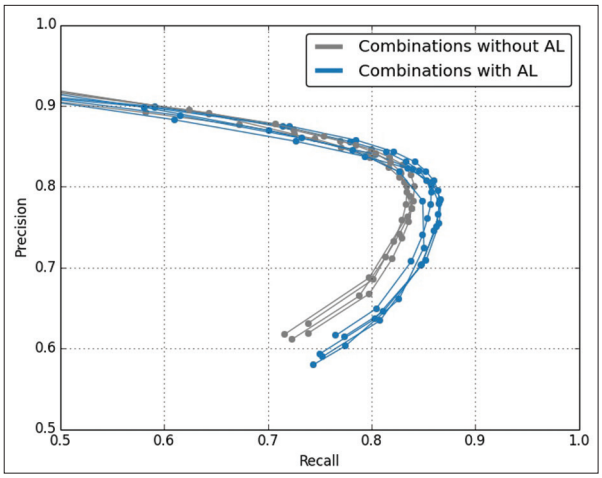

Figure 7: Most approaches that combine multiple selection methods and contain active learning (blue) yield better quality measures than others (gray), here shown for the approaches using Voronoi-based sample extraction

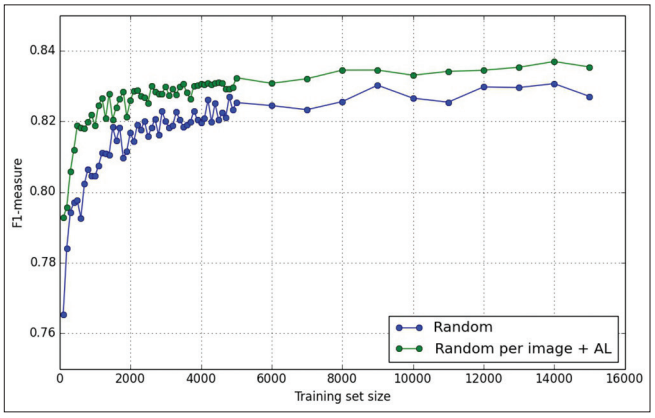

Figure 8: In experiment 3, training sets of different sizes were produced for the approaches described in experiment (1b) and (1e). The graph plots the f1-measures of the nuclei detection trained with these sets against their size

experiment was repeated five times to average out randomness in the used algorithms.

Table 2 shows the results of experiment 4. It appears that the manual extraction yields only marginally better results than the approaches based on center markers. Since the quality of the segmentation-based samples is certainly better than of those of the proposed methods, the low difference of the methods might seem unexpected. However, the produced training sets are all subject to the same selection process so that these quality differences manifest less noticeable in the final results.

Conclusions

The quality of machine learning-based nuclei detection methods is fundamentally dependent on the training data used. Ideally, training data should be generated from manual segmentations of a large number of nuclei, which is, however, a time-consuming and tedious task.

In this paper, we proposed and compared approaches to produce training sets from easy to generate center marker annotations. We divided the training set generation into a sample extraction and a sample selection step. The samples were extracted using a distance-based or a Voronoi boundary-based method. Training sets were selected from the resulting sample sets using different combinations of stratified random subsampling, Kd-tree subsampling, and active learning.

For evaluation, we trained a nuclei detection method with these training sets and assessed the resulting detection quality measures. In addition, we investigated the influence of the cutoff value on these measures. For a cutoff value of 0.5, the default threshold for two-class problems, class balancing had the largest positive impact on the detection quality. Independent of the cutoff value, the best results were obtained using training sets produced by Voronoi-based sample extraction and sample selection methods that incorporate active learning. We also evaluated the influence of the training set size on the detection quality. The quality increased quickly until approximately 2000 samples and more moderate afterward. In a fourth evaluation, a comparison revealed that the f1-measures obtained using the proposed extraction methods almost reached the values obtained using samples generated from manual segmentations.

We conclude that the usage of center marker annotations in conjunction with appropriate sample extraction and selection methods represents a valid alternative to conventionally produced training sets. In this manner, the effort for the creation of annotations can be greatly reduced.

In addition, while machine learning-based nuclei detection methods are usually trained on all training samples available, our study shows that subselecting samples can improve the detection quality considerably at no additional cost in terms of execution time or complexity of the nuclei detection method.

In future work, we will further evaluate the general applicability of the proposed approaches using different image datasets and detection algorithms, especially within the area of deep learning.

Table 2: Quality measures obtained by different extraction methods

Extraction	Precision	Recall	F1-measure
Distance-based	0.791	0.818	0.804
Voronoi-based	0.783	0.831	0.806
Manual segmentation	0.787	0.829	0.807

Financial support and sponsorship

This work was financially supported by Sectra AB, Linköping, Sweden. Part of this work was conducted under the QuantMed project funded by the Fraunhofer Society, Munich, Germany.

Conflicts of interest

There are no conflicts of interest.

References

1. Mahmoud SM, Paish EC, Powe DG, Macmillan RD, Grainge MJ, Lee AH, *et al.* Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast cancer. J Clin Oncol 2011;29:1949-55.
2. Speirs V, Walker RA. New perspectives into the biological and clinical relevance of oestrogen receptors in the human breast. J Pathol 2007;211:499-506.
3. Tang LH, Gonen M, Hedvat C, Modlin IM, Klimstra DS. Objective quantification of the Ki67 proliferative index in neuroendocrine tumors of the gastroenteropancreatic system: A comparison of digital image analysis with manual methods. Am J Surg Pathol 2012;36:1761-70.
4. Vink JP, Van Leeuwen MB, Van Deurzen CH, De Haan G. Efficient nucleus detector in histopathology images. J Microsc 2013;249:124-35.
5. Arteta C, Lempitsky V, Noble JA, Zisserman A. Learning to detect cells using non-overlapping extremal regions. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2012. Berlin, Heidelberg: Springer; 2012. p. 348-56.
6. Kårsnäs A, Dahl AL, Larsen R. Learning histopathological patterns. J Pathol Inform 2011;2:S12.
7. Xing F, Xie Y, Yang L. An automatic learning-based framework for robust nucleus segmentation. IEEE Trans Med Imaging 2016;35:550-66.
8. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 2016;191:214-23.
9. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29.
10. Gul-Mohammed J, Arganda-Carreras I, Andrey P, Galy V, Boudier T. A generic classification-based method for segmentation of nuclei in 3D images of early embryos. BMC Bioinformatics 2014;15:9.
11. Sirinukunwattana K, Raza S, Tsang YW, Snead D, Cree I, Rajpoot N. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans Med Imaging 2016;35:1196-206.
12. Chawla NV, Japkowicz N, Kotcz A. Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explor Newsl 2004;6:1-6.

13. Pechenizkiy M, Puuronen S, Tsymbal A. The impact of sample reduction on PCA-based feature extraction for supervised learning. In: Proceedings of the 2006 ACM Symposium on Applied Computing (SAC). ACM: New York, USA; 2006. p. 553-8.
14. Bentley JL. Multidimensional binary search trees used for associative searching. Commun ACM 1975;18:509-17.
15. Omohundro SM. Efficient algorithms with neural network behavior. Complex Syst 1987;1:273-347.
16. Settles B. Active Learning Literature Survey. Computer Sciences Technical Report 1648. University of Wisconsin-Madison: Madison, USA; 2009.
17. Molin J, Bodén A, Treanor D, Fjeld M, Lundström C. Scale Stain: Multi-Resolution Feature Enhancement in Pathology Visualization, ArXiv Prepr. arXiv:1610.04141; 2016.
18. Kost H, Homeyer A, Bult P, Balkenhol MC, van der Laak JA, Hahn HK. A generic nuclei detection method for histopathological breast images. In: Medical Imaging 2016: Digital Pathology. Bellingham; Washington USA; 2016;9791:97911E-1 - 97911E-7.