Identifying Histological Elements with Convolutional Neural Networks

Christopher Malon
NEC Laboratories America
4 Independence Way
Princeton, NJ 08540
malon (at) nec-labs.com

Matthew Miller
NEC Laboratories America
4 Independence Way
Princeton, NJ 08540
mlm (at) nec-labs.com

Harold Christopher Burger
NEC Laboratories America
4 Independence Way
Princeton, NJ 08540
burger (at) nec-labs.com

Eric Cosatto
NEC Laboratories America
4 Independence Way
Princeton, NJ 08540
cosatto (at) nec-labs.com

Hans Peter Graf
NEC Laboratories America
4 Independence Way
Princeton, NJ 08540
hpg (at) nec-labs.com

ABSTRACT

Histological analysis on stained biopsy samples requires recognizing many kinds of local and structural details, with some awareness of context. Machine learning algorithms such as convolutional networks can be powerful tools for such problems, but often there may not be enough training data to exploit them to their full potential. In this paper, we show how convolutional networks can be combined with appropriate image analysis to achieve high accuracies on three very different tasks in breast and gastric cancer grading, despite the challenge of limited training data. The three problems are to count mitotic figures in the breast, to recognize epithelial layers in the stomach, and to detect signet ring cells.

Categories and Subject Descriptors

I.5.4 [**Pattern recognition**]: Applications—*Computer vision*; J.3 [**Life and medical sciences**]: Medical information systems

General Terms

Algorithms, experimentation

Keywords

Computer-aided diagnosis (CAD), convolutional neural networks (CNN), medical imaging, cancer, oncology, biopsy, histological analysis

1. DIGITAL PATHOLOGY

As pathology data become digitized, there are new opportunities to improve the efficiency and quality of diagnosis. Several studies [13, 16] have demonstrated low agreement among pathologists' grading of the same cases of carcinoma, calling into question the objectivity of diagnoses made by humans alone. Computerized analysis could help pathologists achieve more reproducible results more quickly.

Radiology already has achieved these benefits from digitization [11]. Digital pathology is motivated partly by the success of digital radiology, but faces serious obstacles. Images for pathology are bigger than in radiology, and practical scanners are only now available. Automated analysis of these images will be much more complex, depending on the type of tissue presented and the kind of analysis needed. In a typical pathology problem, a rich variety of tissue is visible, objects overlap within two–dimensional slices, and correct analysis requires aggregating results of many object and structural recognition tasks.

We are developing digital pathology systems targeting breast and gastric cancer. Our system grades digital images of biopsy samples, stained by hematoxylin and eosin. Each slide can be viewed at up to 400X (4.390 dots per micron) power magnification. To reduce computation time, we perform low–resolution analysis at 100X magnification, and select up to eight 256×256 regions of interest to analyze at high resolution (400X magnification).

The Nottingham–Bloom–Richardson score for breast cancer requires a grading of pleomorphism, an estimate of tubule formation, and a count of mitotic cells [6]. For gastric cancer, the shape and distribution of cell nuclei weighs heavily in diagnosis. [7] In all of these tasks, the pathological significance of recognized objects depends upon their context in the tissue. For example, a high density of cell nuclei might be alarming inside stroma, but not inside an epithelial layer.

Vision problems vary in the degree to which they can be solved by pure machine learning and by hand–coded rules. Pure machine learning will work best if sufficient training data is available. However, histological data is very expensive to annotate. Hand–coded rules require no training data, but may be difficult to write and need redesign for every task.

Our approach is to have machine learning perform as much

CSTST 2008 October 27-31, 2008, Cergy-Pontoise, France

of the classification as possible, but to filter candidates by preprocessing that embeds our prior knowledge, until the recognition problem is simple enough to learn from the training data available to us. For our first problem, epithelial layer detection, a convolutional neural network (CNN) produces acceptable results by itself. Data for mitotic figures, our second problem, is scarcer, and more difficult to identify, so we introduce a second trained classifier (using support vector regression) to restrict candidate nuclei to appropriately colored blobs. Signet ring cells, our third identification problem, are difficult even for human pathologists to identify, and they mark a fairly rare form of gastric carcinoma. We restrict candidates for the signet ring cell CNN both with a geometric heuristic and with a second CNN trained for nuclear shape.

2. CONVOLUTIONAL NETWORKS

Traditional neural networks can be overwhelmed by bitmap images. When there are too many unconstrained weights in a neural network, the capacity of the network explodes, and gradient descent does not approach a global minimum loss. Applications of neural networks in digital radiology commonly extract features from the bitmaps before applying the networks [3, 20].

Convolutional neural networks (CNN) [10], however, can be applied succesfully to bitmaps, as they impose equality constraints on many of the weights to simplify the loss minimization problem. In a CNN, one thinks of connections between two–dimensional tensors rather than connections between scalar values. Maps between hidden layers of a CNN implement *convolution* or *subsampling*; in either case, small kernels are convolved with *receptor fields* which sample inputs from each layer. Over the entire CNN, the set of inputs used to determine one output is known as an *input frame*. Layers of the network near the input often represent densities or edge detectors once they are trained. But the designers do not have to decide that these features are necessary; the CNN learns to use them automatically.

Among supervised learning techniques, convolutional neural networks achieve among the highest accuracies on benchmarks such as handwritten digit classification (MNIST, 0.8% error) [10]. But they are particularly attractive choices for time-critical industrial applications in which objects must not only be identified "in frame" but located within a bigger image, because the computation needed to classify overlapping frames can be shared in a natural way. This makes them popular in problems such as face detection [8, 14]. We are aware of just one application of CNN to digital radiology to lung nodules [12].

Our CNN have two outputs, labeled as $\delta_0 = (1, 0)$ and $\delta_1 = (0, 1)$ for negative and positive training examples. We train them using the software package Torch 5 [2] to minimize the loss function

$$L(\vec{x}, \delta_i) = -\log \frac{e^{x_i}}{\sum_j e^{x_j}}$$

so that outputs of the neural network represent log likelihoods of class membership. Training follows the Stochastic Gradient Descent algorithm, a method of backpropagation that often converges faster than batch learning.

Before settling on the various CNN described below, we considered dozens of architectures, with different depths or numbers of units in the hidden layers. Shallower networks with fewer hidden units generally are less susceptible to overfitting, require less training data, and train faster per example. On the other hand, a deeper network with more hidden units may be able to learn the form of the training data more precisely. The series of "LeNet" architectures [10] have been prototypes for many successful applications in image processing, particularly handwriting recognition and face detection. Each of the CNN we describe below is loosely patterned after LeNet 5. Namely, each alternates subsampling layers and 5×5 convolutional layers, with maximum overlap between receptor fields. Among different architectures and learning rates, the best was selected by performance on a hold–out validation set.

A major limitation in training these CNN is the scarcity of training data. Depending on the invariances of a recognition problem, artificial samples may be used to supplement a training set. In each of the problems here, we make use of rotational and reflectional invariance, and supplement each image with its rotations in multiples of 90 degrees, and the corresponding reflections, for a total of eight training inputs per original example.

3. THREE APPLICATIONS

3.1 Epithelial layer detection

Healthy epithelial tissue generally has very different characterics from surrounding tissue, necessitating different processing from other areas. For example, epithelial nuclei are often so large and dense that similar nuclei would indicate malignancy if found in other places. Goblet cells, which are common in epithelia, can resemble signet rings, a serious sign of malignancy in gastric cancer (discussed below). And identifying epithelial tissue is essential for performing certain tasks such as detecting duct formation. For these reasons, it is important to determine which parts of a tissue sample are likely to be epithelial layers before proceeding with other analysis.

Ramírez-Niño, Flores, and Castaño [15] detect epithelia in the context of cervical cancer. Their system uses a linear classifier to classify each pixel into one of four types based on its color. The epithelial boundaries are then determined heuristically based on local histograms of these four pixel types. Tabesh *et. al.* [17] analyze images of prostate tissue, beginning with color segmentation of the image. Epithelial nuclei are identified as leftover nuclei after stromal and apoptotic nuclei are identified based on color, shape, and the classification of surrounding segments. In both cases, we expect the heuristics not to be applicable to epithelial layers in other types of organs.

Epithelial tissue is easily recognized at low resolution, looking only at concentrations of hematoxylin and eosin dye. Accordingly, our CNN for epithelial layer detection processes the image at 31.25X (0.343 dots per micron), after separating the RGB colors into two channels representing the two dyes, as described in [4]. The input consists of 48×48 frames in these two channels. The separation into hematoxylin and eosin is much more efficient than processing a third input plane with the CNN.

For training and testing, we hand-draw binary masks over the epithelial layers of several images, using a paint program. Frames randomly chosen from inside the masked area become positive training examples, and frames randomly chosen from outside become negative examples. Because the

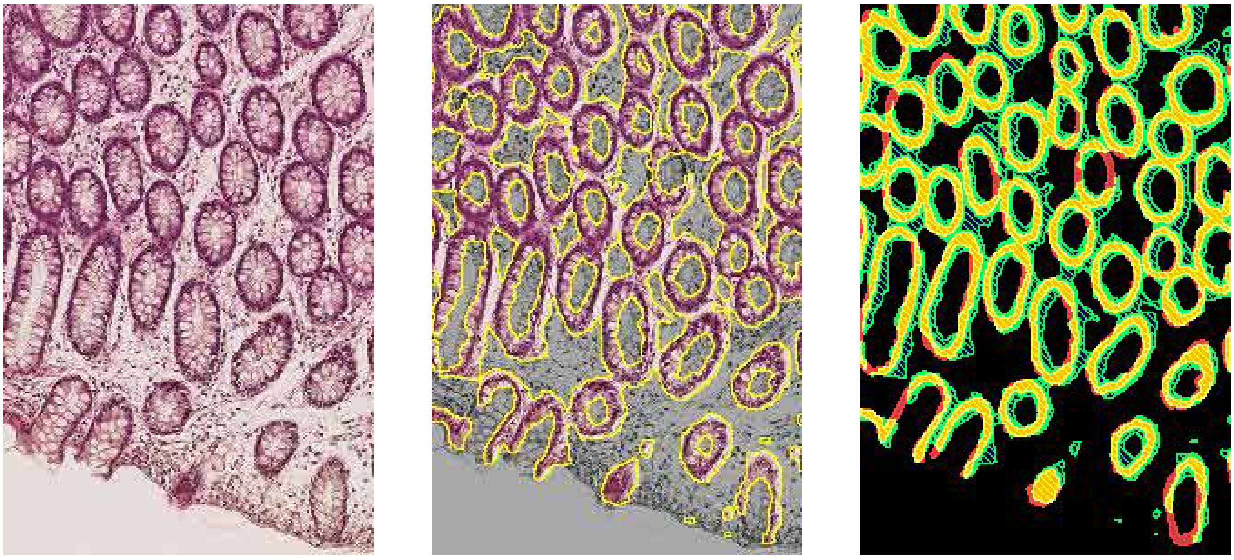

Figure 1: Epithelial layer detection. ***Left: original image, Center: detections highlighted (full brightness against dimmer background), Right: true negative (black), false negative (dark, solid), false positive (dark, hatched), true positive (bright)***

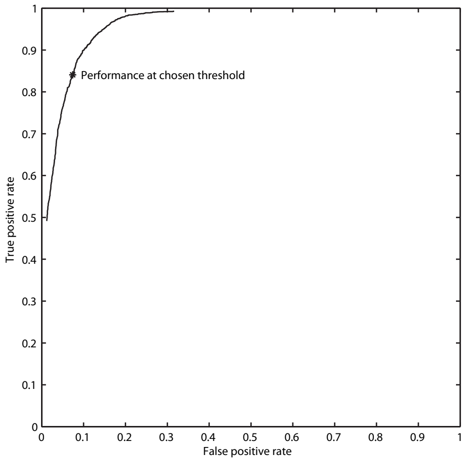

Figure 2: ROC for epithelial layer classification

appearance and pattern of epithelial layers can vary tremendously among patients, it is important to test the system on tissues taken from different patients than the training data. Altogether, there are 21 tissues for training and 4 for testing, producing 9,450 training and 800 testing examples, times 8 for 90-degree rotations and reflections.

An example of the epithelial layer detector's output on a test image is shown in Figure 1. Agreement between the hand–drawn mask (which is somewhat subjective) and computer–generated mask is very good. The ROC curve for frame classification is shown in Figure 2. At the chosen threshold, the detector produces 7.4% false positives and 84.1% true positives on the test set.

3.2 Mitosis

Mitotic count is one of three criteria (along with pleomorphism and tubularity) used to compute the Nottingham–Bloom–Richardson grade [6]. We are developing a system to reproduce this grading in its three components. Our work on the pleomorphism grade is described in [4]. Here, we describe how we count the mitotic figures.

Mitosis follows four phases—prophase, metaphase, anaphase, and telophase—but we train one classifier that recognizes any of them simply as "mitosis." Mitosis can only be detected at high resolution, so we train our classifier at 400X magnification. The training and validation data comes from a set of 728 images, 1024 by 768 pixels at this resolution, on which a pathologist searched for all mitotic figures and identified 434.[1] Of this set, 65% is used for training of classifiers and 35% is used as a hold–out validation set.

To train an effective CNN with so few positive examples, the negative examples must be as instructive as possible. Most regions of the tissue will not have any mitotic figures at all, and can be eliminated heuristically. Observing that mitotic nuclei exhibit discoloration compared to normal nuclei, we define candidates for the CNN to be sufficiently large blobs of points that meet some color criteria.

These color criteria may be defined in more or less naive ways. Each of them utilizes the color histogram of all nuclei in the input image, which we can determine *a priori* because nuclei are marked by having exceptionally high hematoxylin content. The most naive approach we consider is a "color box," in which we pre–assign permissible color ranges in the red, green, and blue channels. These ranges are chosen to have pre–determined widths around the peaks of the color histogram in each channel. Nuclei whose colors fall within the boxes are regarded as non–mitotic. A slightly stronger approach, referred to in Figure 3 as "color histograms," adjusts the ranges so the underlying integrals of the color histogram over the ranges equal some constant. But the

[1] A second pathologist was given the same problem, and identified 515 figures, 271 of them in common.

strongest method is to train a classifier by support vector regression (SVR) [18] to predict the mitotic color thresholds, from the overall nuclear color histogram of the image. At the parameters marked with an arrow in Figure 3, this method misses just 22 figures (10.5%) while producing only 6,904 candidates on the validation set. An example of the candidate detections is shown in Figure 4.

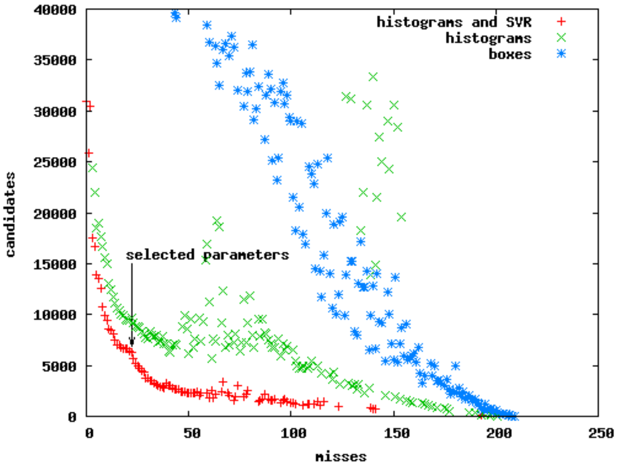

Figure 3: Mitosis candidate selection

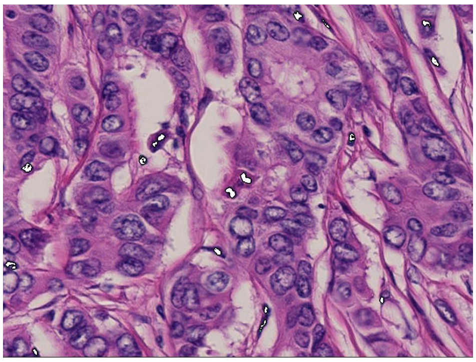

Figure 4: Candidates for the mitosis detector (bright white)

The CNN operates on the candidates selected by the SVR color preprocessing. Input tensors consist of the red, green, and blue channels, in a 60 by 60 (at 400X) pixel frame around the center of the candidate figure. Because of the overwhelming abundance of negative examples, positive examples must be promoted in the stochastic gradient descent algorithm, over their natural appearance in the data set. We do this by drawing one positive example for every five negative examples, in our random example selection.

An ROC curve may be obtained by varying a threshold for the difference in CNN outputs. Figure 5 shows that generalization is good, as results on the training and validation sets are close. One may obtain 80% of positives for a false positive rate of about 5%. Other systems to find mitotic nuclei on stained images have reported recall rates of 92–95% for 22–42% false positive [1]. As the only negative samples classified by our CNN are those that pass our rigorous dis-

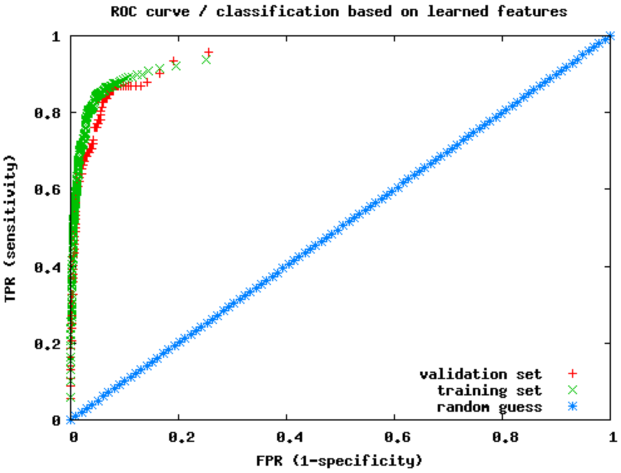

Figure 5: ROC for mitosis detector CNN

coloration criteria, one can expect that our system is much stronger.

To compute the mitotic component of the Nottingham–Bloom–Richardson score, we have integrated this system into a module in conjunction with an SVM to predict one of the three mitotic grades defined by this system, on an entire tissue. On our test set, our system's grade agreement with a pathologist on this three–class problem is $\kappa = .40$. This performance approaches agreement between two pathologists on mitotic grading, which is reported to be in the range $\kappa = .45$ to $\kappa = .64$ [13]. We have not seen other systems that attempt to compute a mitotic grade on a whole tissue.

3.3 Signet ring cell detection

One kind of gastric cancer is indicated primarily by the appearance of *signet ring cells* within the tissue. These cells are distinguished by the appearance of squashed nuclei on their periphery. Typically, they have mucinous cytoplasm and do not form glands or tubules [7]. In contrast to other kinds of gastric cancer, which exhibit histological phenomena across a large tissue region, signet ring cells may occur only at a few dozen sites, and are easily missed in microscopic examination [5, 9].

The first step in signet ring cell detection is to determine candidates by geometric preprocessing. A CNN trained to identify squashed nuclei then eliminates some candidates from this list. Only the candidate cell centers with squashed nuclei nearby remain as candidates for a second CNN, which judges whether the overall cell configuration appears like a signet ring.

The image analysis step, which is looking for potential cell centers, considers the radial symmetry of cell membranes. Because of their eosin content, cell membranes tend to contrast more strongly with surrounding channel response on the green channel than on the others. An edge detector works in horizontal, vertical, and diagonal directions to produce four edge maps on the green channel. Edge responses are enhanced with a nonlinear filter. Then, the candidate region for cell centers is computed. The region excludes areas of the wrong color (particularly, white or blood regions) and points that appear too close to the detected edges.

On the remaining region, a generalized Hough transform is computed from the four edge maps. The form of this transform, applied to a greyscale bitmap B at (x, y), is

$$H(x,y) = \frac{1}{C} \int_0^{\pi} \int_{r_1}^{r_2} f(\quad B(x + r\cos\theta, y + r\sin\theta), \\ B(x - r\cos\theta, y - r\sin\theta)) \quad dr\ d\theta$$

where $f(a, b)$ is one if a and b both exceed a given threshold, and zero otherwise. Thus, $H(x, y)$ measures the radial symmetry about (x, y). We apply a discretized version of this transform, using the four edge maps in each $\pi/4$ interval of the integral. Candidate points for signet ring cells are selected as the peaks of this transform achieving a given threshold, provided that no two peaks are chosen too close together.

These candidate points are pruned using a CNN trained as a squashed nucleus detector. As in the mitosis detector, nuclei may be identified as regions of high hematoxylin content. The input to the detector consists of 48×48 binary images at 400X magnification, centered in the middle of the candidate squashed nuclei, representing nuclear shape alone.

When a Hough peak occurs near a squashed nucleus, a second CNN judges the overall configuration of the tissue around the candidate, to make the final judgment of whether or not a signet ring cell is present. It utilizes red, green, and blue color channels, in a 204×204 (at 400X) frame about the Hough peak. As this CNN incorporates color information and judges more than shape, it requires more capacity than the CNN for squashed nuclei.

Cases of signet ring cell cancer are fairly rare, even within gastric carcinoma. In a gastric dataset of 2,328 tissue samples from 896 patients, only 29 tissues, from 10 patients, were positive for signet ring cell cancer. Because we needed to obtain training, validation, and testing sets from this set, each from disjoint sets of patients, our hold–out validation set included positive examples from only two patients. Two pathologists selected the cases of signet ring cell cancer, but we labeled the locations of individual signet rings ourselves. Negative examples came from randomly chosen Hough peaks on tissues without signet ring cell carcinoma. We trained the system on a balanced data set of 4,022 examples, plus their rotations and reflections. Additionally, we marked 626 examples for validation and 2,650 for testing.

Training examples for the squashed nucleus detector also were selected by hand. From the nuclei surrounding each ground truth signet cell, we marked ones that appeared visibly squashed. Negative squashed nuclei examples were chosen at random from nuclei near the Hough peaks selected as non–signet ring examples.

The example in Figure 8 illustrates each step in the classification: squashed nucleus detection, the search for nearby Hough peaks, and the classification of those peaks. Figure 6 presents the ROC curves for both CNN. Deciding the signet ring cell configuration appears to be a difficult problem, more challenging than the determination of squashed nuclei. As the gap between training and validation performance suggests, there is not enough training data for good generalization. Therefore, it is beneficial to add the specialized CNN for squashed nuclei to avoid false positives. The improvement in tissue classification that is achieved when the two CNN are used together is shown in Figure 7.

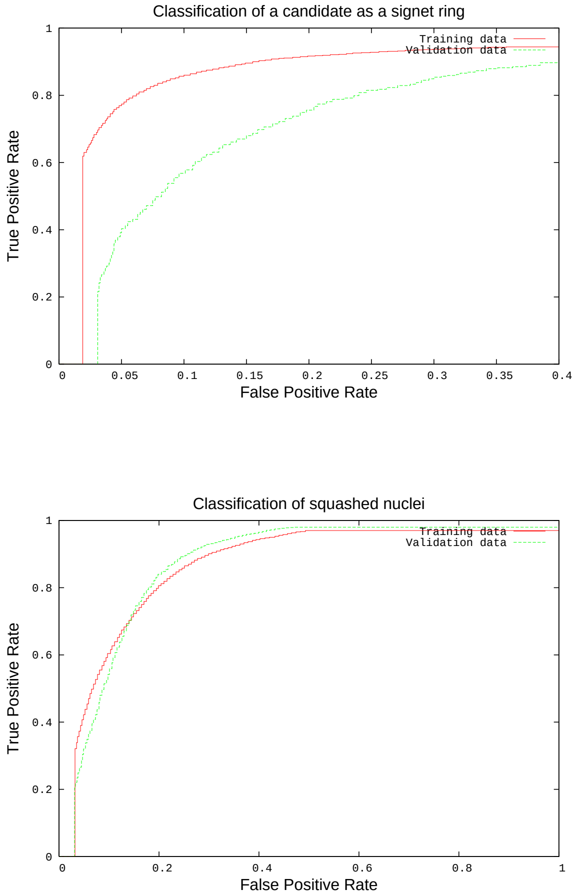

Figure 6: ROC for each signet ring CNN

We do not have recall and alarm rates for a human pathologist, but cases missed by three pathologists are not uncommon [9], and some literature suggests that a new kind of stain is needed to find signet ring cells reliably [5]. We suspect that our detector performs competitively with human pathologists and can expose missed cases by enabling tissue to be scanned more thoroughly.

4. CONCLUSION

We have established convolutional neural networks as a versatile technique for detecting regions of pathological significance in biopsy images. In cases where the amount of training data is insufficient, we have achieved better performance by simplifying the problems for the CNN to learn, with image analysis that restricts the cases for training and classification. It is critical that we do not compute expensive features on entire tissue images. With CNN, much of the computation needed to classify overlapping frames is common and can be performed just once.

The burden of obtaining a large set of data with detailed labels by professionals is a significant obstacle to any supervised machine learning technique applied to medical diagnosis. *Semi-supervised* learning techniques, such as [19], make use of vast pools of unlabeled examples to achieve stronger classification even with relatively little labeled training data. A CNN trained on labeled data can be replaced easily by one trained with semi-supervised learning techniques, and

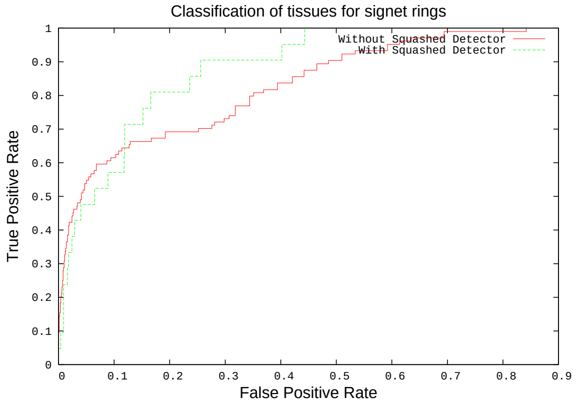

Figure 7: Effect of adding squashed nucleus detector

we will take advantage of this opportunity in future work.

5. ACKNOWLEDGMENTS

We thank Dr. John S. Meyer, M.D., formerly with St. Luke's Hospital, Chesterfield, MO, USA, for consultations regarding grading, and for preparation of our test set of mitotic figures.

6. REFERENCES

[1] J. A. M. Beliën, J. P. A. Baak, P. J. van Diest, and A. H. M. van Ginkel. Counting mitoses by image processing in feulgen stained breast cancer sections: The influence of resolution. *Cytometry*, 28:135–140, 1997.

[2] R. Collobert, S. Bengio, L. Bottou, J. Weston, and I. Melvin. Torch 5, `http://torch5.sourceforge.net`.

[3] G. Coppini, S. Diciotti, M. Falchini, N. Villari, and G. Valli. Neural networks for computed-aided diagnosis: Detection of lung nodules in chest radiograms. *IEEE Transactions on Information Technology in Biomedicine*, 7(4), December 2005.

[4] E. Cosatto, M. Miller, H. Graf, and J. Meyer. Grading nuclear pleomorphism on histological micrographs. In *19th International Conference on Pattern Recognition*, 2008. To appear.

[5] H. M. T. El-Zimaity, K. Itani, and D. Y. Graham. Early diagnosis of signet ring cell carcinoma of the stomach: role of the genta stain. *J. Clin. Pathol.*, 50:867–868, 1997.

[6] C. W. Elston and I. O. Ellis. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19:403–410, 1991.

[7] J. R. S. for Gastric Cancer. *Japanese Classification of Gastric Carcinoma*. Kanehara & Co., Ltd., 1995.

[8] C. Garcia and M. Delakis. Convolutional face finder: A neural architecture for fast and robust face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 2004.

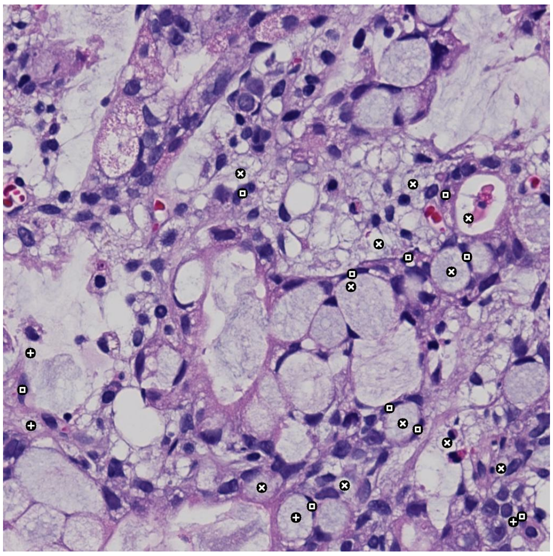

Figure 8: Signet ring cell classification. *Square: squashed nucleus; Plus: negative Hough peak, Cross: positive Hough peak (center of signet ring).*

[9] R. Gupta, R. Arora, P. Das, and M. K. Singh. Deeply eosiniphilic cell variant of signet–ring type of gastric carcinoma: a diagnostic dilemma. *Int. J. Clin. Oncol.*, 13:181–184, 2008.

[10] Y. Le Cun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[11] J. Lee. Medical technology update - a Canadian perspective. *Canadian Journal of Medical Radiation Technology*, 36(3):26–33, 2005.

[12] S.-C. B. Lo, S.-L. A. Lou, J.-S. Lin, M. T. Freedman, and M. V. Chien. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, 14(4), December 1995.

[13] J. S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, A. Glass, B. A. Zehnbauer, K. Lister, and R. Parwaresch. Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. *Modern Pathology*, 18:1067–1078, 2005.

[14] M. Osadchy, Y. Le Cun, and M. L. Miller. Synergistic face detection and pose estimation with energy-based models. *J. Machine Learning Research*, 8, 2007.

[15] J. Ramírez-Niño, M. A. Flores, and V. M. Castaño. Image processing and neural networks for early detection of histological changes. In R. M. et al., editor, *MICAI 2004: Advances in Artificial Intelligence (LNAI 2972)*, pages 632–641. Springer, 2004.

[16] P. Robbins, S. Pinder, and N. de Klerk. Histological grading of breast carcinomas: A study of interobserver agreement. *Hum. Pathol.*, 26:873–876, 1995.

[17] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi. Multifeature

prostate cancer diagnosis and gleason grading of histological images. *IEEE Transactions on Medical Imaging*, 26(10):1366–1378, October 2007.

[18] V. Vapnik. *Statistical Learning Theory, Second edition*. John Wiley & Sons, 1998.

[19] J. Weston, F. Ratle, and R. Collobert. Deep learning via semi-supervised embedding. In A. McCallum and S. Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 160–167. Omnipress, 2008.

[20] H.-Y. M. Yeh, Y.-M. F. Lure, and J.-S. Lin. Method and system for the detection of lung nodule in radiological images using digital image processing and artificial neural network, U.S. Patent 6,760,468, 2004.