

# Automated Detection of Diagnostically Relevant Regions in H&E Stained Digital Pathology Slides

Claus Bahlmann<sup>1</sup>, Amar Patel<sup>1</sup>, Jeffrey Johnson<sup>1</sup>, Jie Ni<sup>2\*</sup>, Andrei Chekkoury<sup>1</sup>, Parmeshwar Khurd<sup>1</sup>,  
Ali Kamen<sup>1</sup>, Leo Grady<sup>1</sup>, Elizabeth Krupinski<sup>3</sup>, Anna Graham<sup>3</sup>, Ronald Weinstein<sup>3</sup>

<sup>1</sup>Siemens Corporate Research, 755 College Road East, Princeton NJ, 08540, United States

<sup>2</sup>University of Maryland, MD 20742, United States

<sup>3</sup>The University of Arizona, Tucson AZ 85721, United States

## ABSTRACT

We present a computationally efficient method for analyzing H&E stained digital pathology slides with the objective of discriminating diagnostically relevant vs. irrelevant regions. Such technology is useful for several applications: (1) It can speed up computer aided diagnosis (CAD) for histopathology based cancer detection and grading by an order of magnitude through a triage-like preprocessing and pruning. (2) It can improve the response time for an interactive digital pathology workstation (which is usually dealing with several GByte digital pathology slides), e.g., through controlling adaptive compression or prioritization algorithms. (3) It can support the detection and grading workflow for expert pathologists in a semi-automated diagnosis, hereby increasing throughput and accuracy. At the core of the presented method is the statistical characterization of tissue components that are indicative for the pathologist's decision about malignancy vs. benignity, such as, nuclei, tubules, cytoplasm, etc. In order to allow for effective yet computationally efficient processing, we propose visual descriptors that capture the distribution of color intensities observed for nuclei and cytoplasm. Discrimination between statistics of relevant vs. irrelevant regions is learned from annotated data, and inference is performed via linear classification. We validate the proposed method both qualitatively and quantitatively. Experiments show a cross validation error rate of 1.4%. We further show that the proposed method can prune  $\approx 90\%$  of the area of pathological slides while maintaining 100% of all relevant information, which allows for a speedup of a factor of 10 for CAD systems.

**Keywords:** Breast histopathology, high-speed CAD histology, breast cancer, cancer detection from digital pathology, triaging & pruning

## 1. DESCRIPTION OF PURPOSE

This work deals with virtual slides from H&E (hematoxylin & eosin) stained digital histopathology, such as illustrated in Figure 1. Such slides are usually several GByte in size and their analysis by pathologists and computer algorithms is often limited by the technologies currently available for digital pathology workstations.<sup>1</sup> We present a method that aims at facilitating and accelerating the analysis of virtual slides by automatically identifying diagnostically relevant regions in such slides, and at the same time discarding most of the irrelevant ones. Such functionality can be beneficial in a number of applications, e.g.:

1. It can speed up computer aided diagnosis (CAD) for histopathology based cancer detection and grading<sup>2-7</sup> by an order of magnitude through a triage-like preprocessing.
2. It can improve the response time for an interactive digital pathology workstation dealing with several GByte large virtual slides, e.g., through controlling adaptive compression or prioritization algorithms.<sup>1,8,9</sup>
3. It can support the detection and grading workflow for expert pathologists in a semi-automated diagnosis, thereby increasing throughput and accuracy.

Our algorithm design addresses two main requirements. First, since the algorithm will be used in a triage-like preprocessing context, we aim at almost 100% detection accuracy, while false alarm rate should be low, but not necessarily zero. Second, since our algorithm is usually applied at the entire large virtual slide, e.g., for pruning, computational speed should be high, with further potential improvement using hardware speedup, e.g., cluster or GPU processing.

---

\*work performed while at Siemens Corporate Research

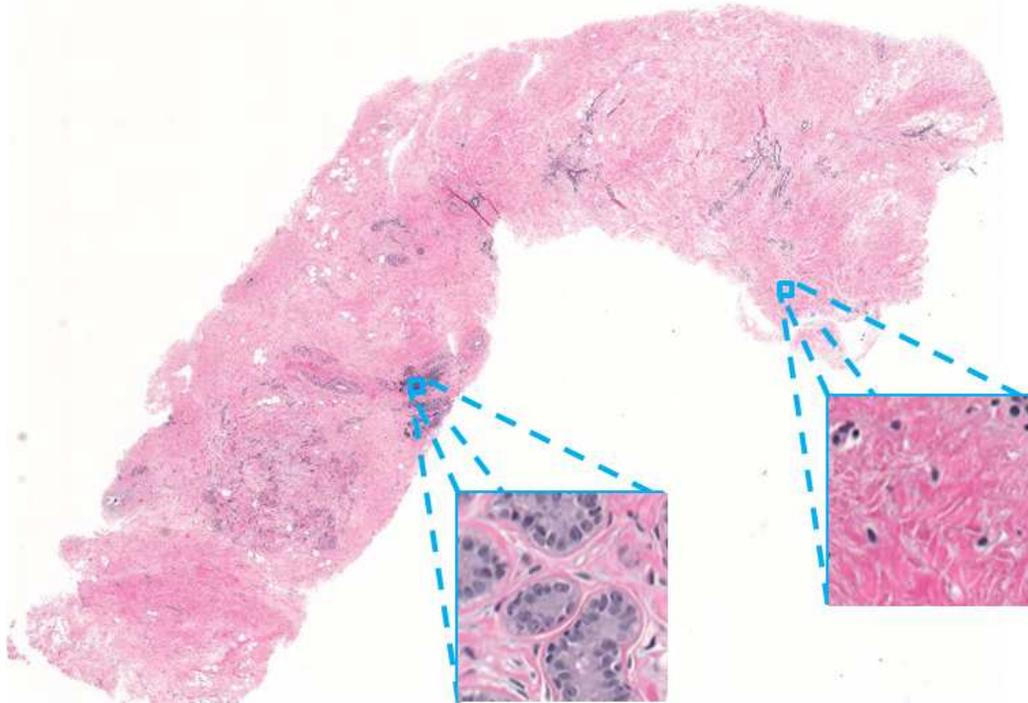


Figure 1: Example of a  $\approx 4$  GByte virtual slide for a breast biopsy specimen with two close-up views of a diagnostically relevant (left) and irrelevant (right) regions. The difference between the two samples can be clearly seen by the number of indicative elements of nuclei, tubules, cytoplasm, etc.

## 2. METHODS

The virtual slides in our study were acquired for breast biopsy specimens using a DMetrix scanner in the Arizona Telemedicine Program. Slide images were sampled at  $0.47 \mu\text{m}/\text{pixel}$ . For a typical slide with 1 to  $4 \text{ cm}^2$  of tissue, a single 40X objective scan yields 1 to 5 GB of uncompressed RGB image data. Figure 1 shows an example of a virtual slide of about  $40000 \times 30000$  pixels resolution.

Two close-up views show examples of different tissue regions that were classified by an expert pathologist as *relevant* or *irrelevant* to the diagnosis of breast cancer.

Diagnostically relevant regions are distinguished by a large amount of epithelial nuclei and tubule formation, whereas irrelevant regions are dominated by cytoplasm tissue. In H&E stained images, these tissue components are stained dark purple (nuclei) and pink (cytoplasm and the extracellular connective tissue).

Pathologists typically start by visually scanning a virtual slide to identify the most diagnostically relevant tissue. Our proposed automated detection follows this procedure by subdividing the slide into square image patches of  $256 \times 256$  pixels (corresponding to  $120 \times 120 \mu\text{m}$ ). Similar to human pathologist processing, it aims at modeling the distribution of nuclei and cytoplasm. Specifically, it employs a combination of color preprocessing, the extraction of feature descriptors, and classification based on machine learning, as is illustrated in Figure 2. In the following we will provide details for the steps involved.

### 2.1 Color representation

In H&E stained specimens, nuclei appear purple and cytoplasm appears pink. The proposed approach is aiming at characterizing distributions of these components. In order to accentuate the difference between these colors, we rely on a (linear) color transform into two channels, called H and E, each of them amplifying the hematoxylin (eosin) stain and at the same time suppressing the eosin (hematoxylin) stain. This approach is similar to the approach reported by Cosatto et al.<sup>2</sup> In a nutshell, it computes dominant purple and non-purple pixel values from the data and subsequently computes the main axes for the transform orthogonal to those. Figure 3 shows the axes of dominant pixel values (in black) and the transformation

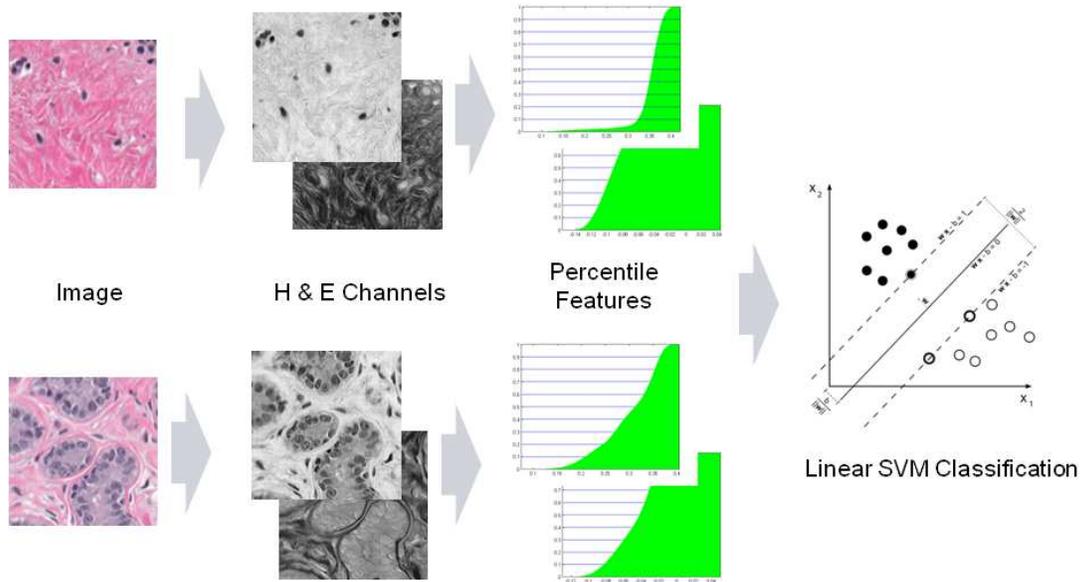


Figure 2: Summary of the proposed approach: An image region is transformed into H and E color channels, percentile feature descriptors are extracted, and classified with a linear SVM classifier.

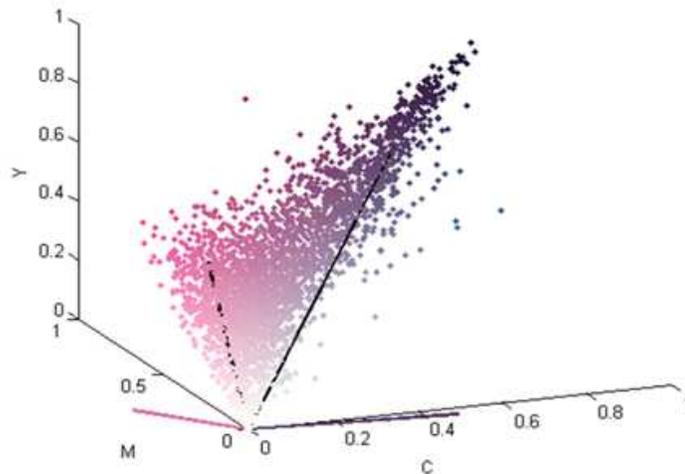


Figure 3: Color transform. The dominant purple and non-purple pixels (illustrated by the two black vectors) are estimated from the data. The linear transformations H and E are computed orthogonal to M and C.

axes (the “lower” axes) for a typical sample set of pixels. For illustration, every point in CMY space has been colored with its particular color.

## 2.2 Descriptor and Classifier

Similar to the process of a human pathologist who considers the distribution of nuclei and cytoplasm within a region, our automated processing is based on the distribution of nuclei pixels and cytoplasm pixels. We have chosen the level of pixels rather than higher abstraction levels, such as shape information, to achieve greater computational speed. In Section 3 we will show that this approximation will already provide sufficient accuracy for the task.

The proposed descriptor is based on the distribution of observed intensities in the pair of H and E channels. Our method chooses a sparse representation of 11 uniformly distributed percentile ranks (at 0%, 10%, 20%, ..., 100%). In practice,

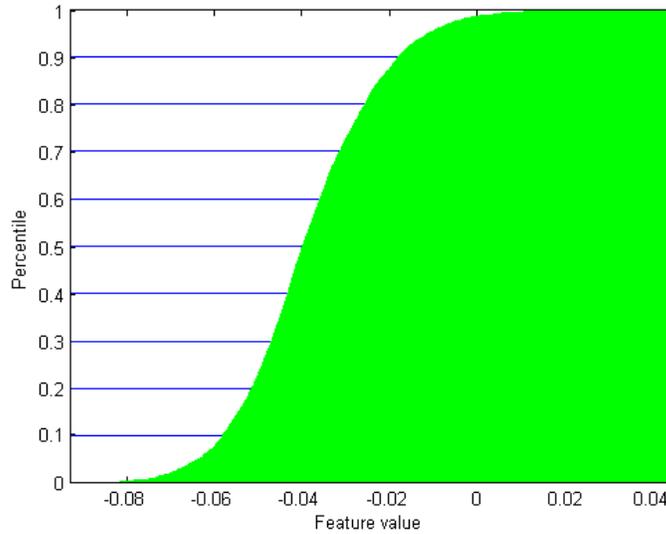


Figure 4: Percentile descriptor

the rank values can be obtained via sorting or by cumulative histogramming, as illustrated in Figure 4. The figure plots the normalized cumulative histogram as a function of intensity (here for the E channel). The descriptor takes values from the abscissa at locations where the cumulative histogram cuts the respective percentile levels.

We also experimented with a different, more common representation that is based on normalized histogram bin counts. The percentile based representation, however, has the benefit that it is adaptive to the range in the feature space, hence, does not require a selection of bin size and range. Notably, with the percentile based representation we empirically observed a relative 50% lower error rate compared to the normalized histogram bin count.

The pair (corresponding to H&E) of 11 percentile values is then combined into a 22 dimensional feature vector, and a linear SVM (using libsvm<sup>10</sup>) is trained for the classification task.

### 3. RESULTS

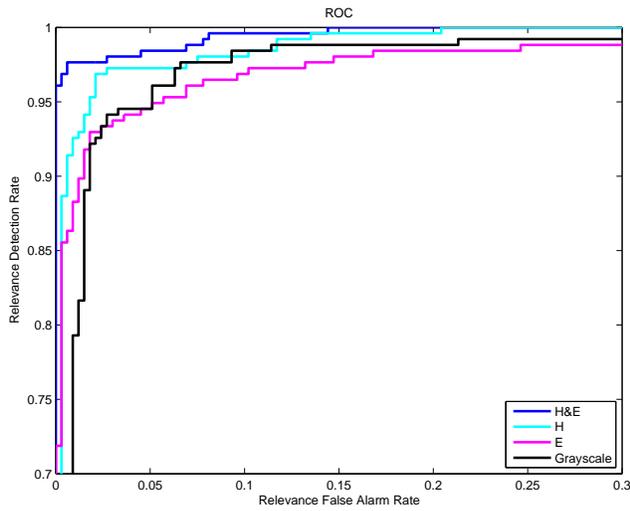
We examine our method with two types of experiments: first, on a set of 589 cropped patches that has been labeled by pathologists as relevant (256 count) or irrelevant (333 count), respectively; second, on full virtual slides of 1-5 GBytes, where pathologists have selectively marked areas of relevance and irrelevance.

#### 3.1 Results on image patches

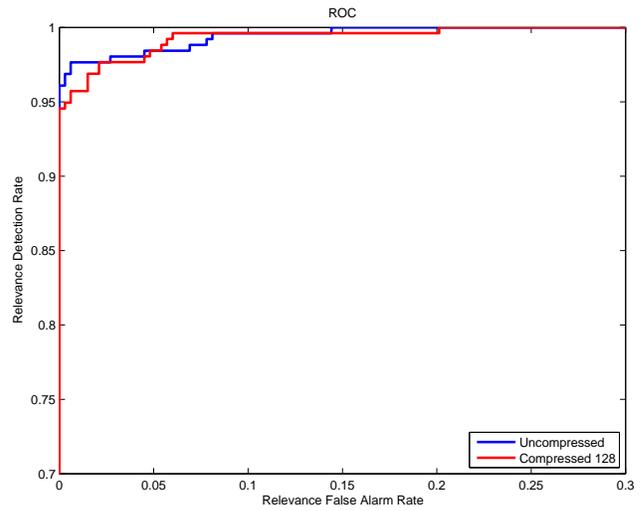
The classification on the cropped patches was evaluated using 50 fold cross validation. Figure 5 (a) shows an ROC curve for this experiment, showing that almost 100% detection rate can be obtained with only 7-8% false positive rate (blue line). The latter is particularly notable, because a CAD based malignancy detection would not see any significant degrade in performance, when combined with the proposed method as a pruning, but would benefit significantly from the speedup. The obtained error rate from the point of the ROC chosen by the SVM is 1.4%.

Figure 5 (a) also compares the obtained performance with different color representations, i.e., grayscale and individual H and E channel, respectively. It shows that the joint H-E color representation gives an advantage over those representations.

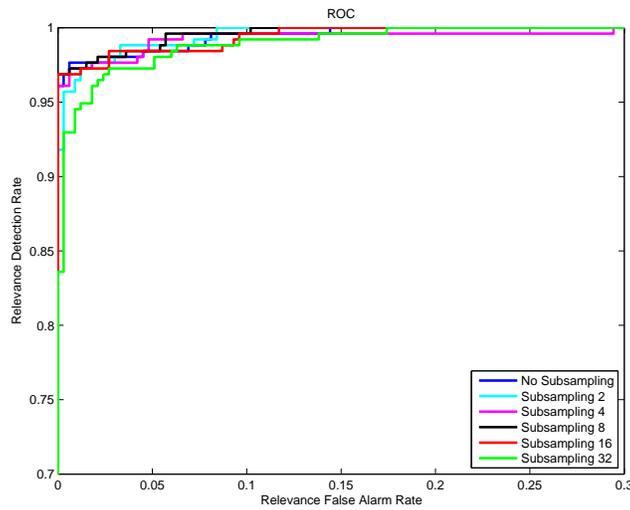
Figure 6 summarizes the failure cases from the experiments with the H&E experiments. From visual verification and comparison to the whole set of images we observed that the misclassifications correspond mostly to outliers or borderline cases.



(a) ROC for different color representations



(b) ROC for different compression levels



(c) ROC for different sub sampling. "Subsampling 16" means sub-sampling on a  $16 \times 16$  grid.

Figure 5: ROC for different pre-processing and feature extraction.

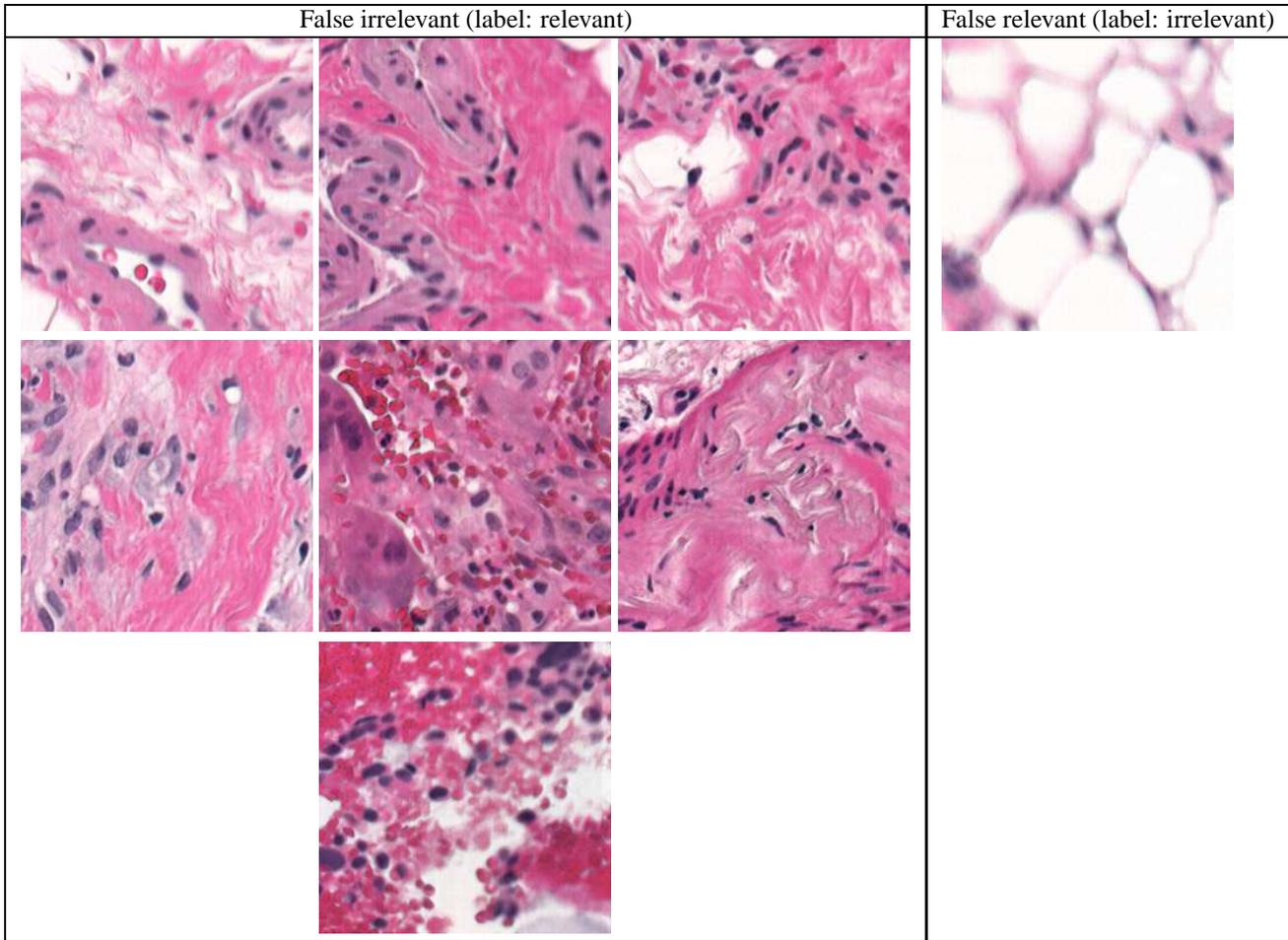


Figure 6: Failure cases ( $8/589 = 1.4\%$ ): to the left those falsely classified as irrelevant, to the right as relevant.

### 3.2 Results on compressed image patches

As image compression plays a major role in digital pathology with large virtual slides, we were further interested in the effect of compression on the presented classification methods. We have evaluated the classification described above with (i) uncompressed images and (ii) images that were JPEG2000-compressed with a factor of 128. Figure 5 (b) illustrates the results. The ROC shows that no significant classification performance is lost from the compression. This robustness to even high compression levels can be explained by the fact that classification mostly depends on the color distribution, which is not very adversely effected by compression.

This result motivated another empirical study in order to quantify the effect of subsampling on the performance. The benefit of such procedure is the reduced computational cost, which — since image operations are solely pixel based — depend directly from the subsampling. From Figure 5 (c) we see that no or only little performance loss can be observed up to subsampling by  $8 \times 8$  or even  $16 \times 16$ , leading to a 64–256 times speedup.

### 3.3 Results on virtual slides

We also examined our method on full virtual slides by scanning individual  $256 \times 256$  pixel patches in a moving window (with zero overlap). Figure 7 shows a fully marked virtual slide where the green regions represent areas that our method classified as diagnostically relevant. The blue and red boxes are non-exhaustively labeled regions that a pathologist annotated as relevant and irrelevant, respectively, meaning that non-annotated regions can belong to either class. We find that in this particular case 100% of the regions annotated as relevant appropriate detections were correctly made, while also all irrelevant regions have been correctly identified. Another interesting note is that  $\sim 90\%$  of the tissue in the slide has been

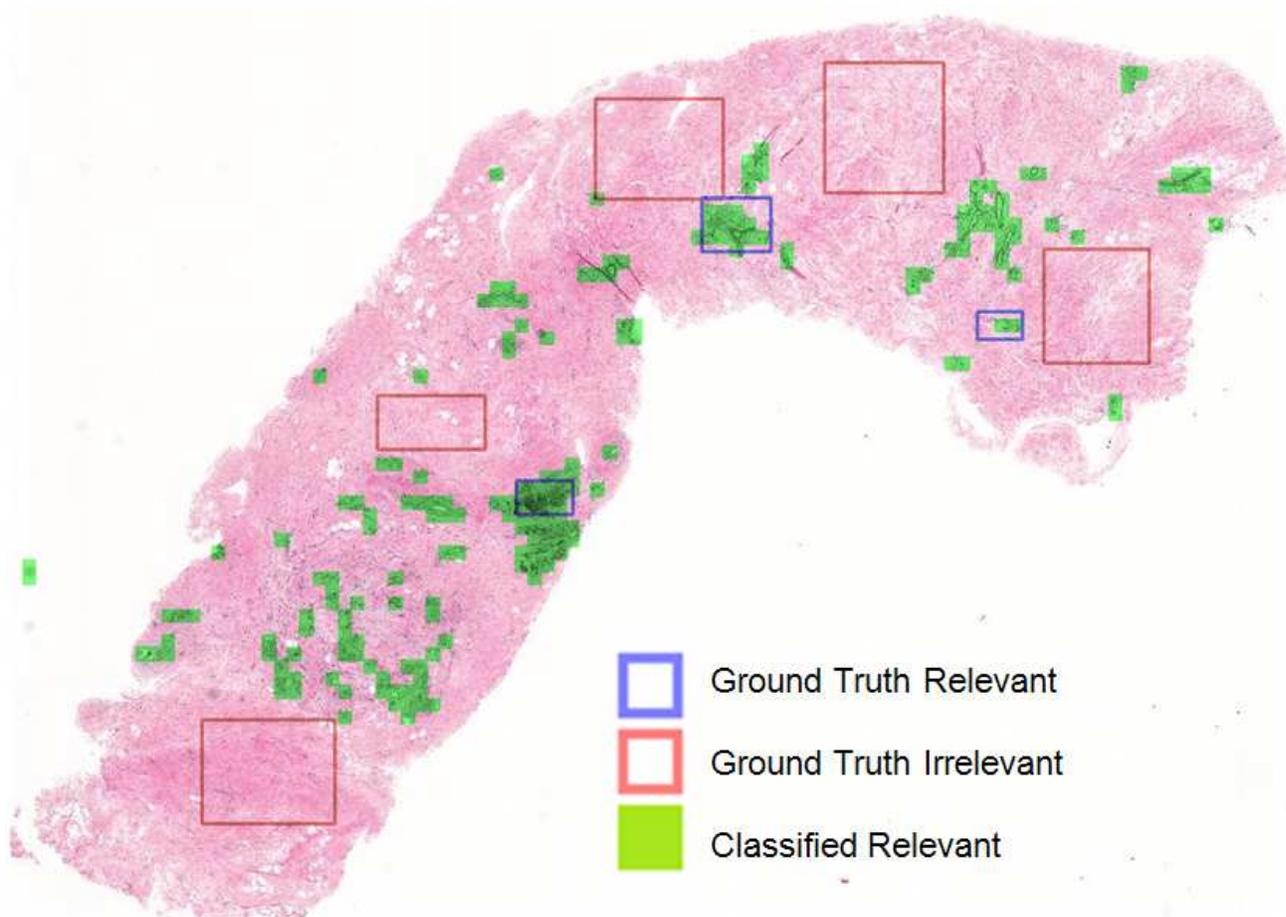


Figure 7: Experiment on a virtual slide.

deemed irrelevant by the automated processing. This indicates a possible system speedup of a CAD system of more than one order of magnitude. Similar observations have been found on 93 additional virtual slides.

### 3.4 Computational speed

Computational speed is currently 0.1 ms for a  $256 \times 256$  patch (i.e.,  $\approx 120\mu\text{m} \times 120\mu\text{m}$ ) on a standard laptop after sub-sampling by  $16 \times 16$ . For a 4 GByte virtual slide, such as the one in Figure 1, processing scales to  $\approx 2$  seconds. In the context of a triaging (or pruning) for CAD systems, this is orders of magnitude faster than popular textron based approaches for histopathology analysis<sup>3</sup> or higher level analysis, hence, it would not increase the overall processing notably.

## 4. CONCLUSION

We have presented an approach for automated image based triaging in digital pathology. The innovation of this work is a computationally efficient algorithm that identifies regions of diagnostic relevance in histopathology virtual slides with high accuracy. This algorithm can serve as a fast triaging or pruning step in a CAD based cancer detection or digital pathology workstations, thereby improving computation and system response time by an order of magnitude. Computational efficiency is achieved by local pixel-based analysis and a sparse color distribution descriptor. Experiments indicate very high accuracy (only 1.4 % error) and up to 10 times speedup potential for the intended application scenarios.

## Acknowledgments

This work was supported by National Institutes of Health/National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB) under Grant R01EB008055.

## REFERENCES

1. E. Patterson, M. Rayo, C. Gill, and M. Gurcan, "Barriers and facilitators to adoption of soft copy interpretation from the user perspective: Lessons learned from filmless radiology for slideless pathology,," *J Pathol Inform.* **2**(1), 2011.
2. E. Cosatto, M. Miller, H. P. Graf, and J. Meyer, "Grading nuclear pleomorphism on histological micrographs,," in *ICPR*, 2008.
3. P. Khurd, C. Bahlmann, P. Maday, A. Kamen, S. Gibbs-Strauss, E. M. Genega, and J. V. Frangioni, "Computer-aided gleason grading of prostate cancer histopathological images using texton forests,," in *Proceedings of the 2010 IEEE international conference on Biomedical imaging: from nano to Macro, ISBI'10*, pp. 636–639, (Piscataway, NJ, USA), 2010.
4. P. Khurd, L. Grady, A. Kamen, S. Gibbs-Strauss, E. Genega, and J. V. Frangioni, "Network cycle features: Application to computer-aided gleason grading of prostate cancer histopathological images,," in *ISBI*, pp. 1632–1636, 2011.
5. S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. D. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology,," in *ISBI*, pp. 284–287, 2008.
6. P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis,," *IEEE Trans. Med. Imaging* **28**(7), pp. 1037–1050, 2009.
7. 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 Trans. Med. Imaging* **26**(10), pp. 1366–1378, 2007.
8. E. Krupinski, "Virtual slide telepathology workstation-of-the-future: lessons learned from teleradiology,," *Sem Diag Path* **26**, pp. 194–205, 2009.
9. J. P. Johnson, E. A. Krupinski, M. Yan, H. Roehrig, A. R. Graham, and R. S. Weinstein, "Using a visual discrimination model for the detection of compression artifacts in virtual pathology images,," *IEEE Trans. Med. Imaging* **30**(2), pp. 306–314, 2011.
10. C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines,," *ACM Transactions on Intelligent Systems and Technology* **2**, pp. 27:1–27:27, 2011.