Title of the article:

The need for careful data collection for pattern recognition in digital pathology

Author:

Raphaël Marée, Dept. EE & CS (Montefiore Institute), University of Liège.

Abstract:

Effective pattern recognition requires carefully designed ground-truth datasets. In this technical note, we first summarize potential data collection issues in digital pathology and then propose guidelines to build more realistic ground-truth datasets and to control their quality. We hope our comments will foster the effective application of pattern recognition approaches in digital pathology.

Key-words:

Pattern Recognition, Object Recognition, Ground-Truth, Quality Control, Digital Pathology

Key Messages (Provide appropriate messages of about 35-50 words to be printed in centre box):

We stress the need for quality control of ground-truth datasets for effective pattern recognition in digital pathology.

Introduction:

In pathology, the study of cells (cytology) and tissue (histology) is performed by examining cells and tissues which were sectioned, stained and mounted on a microscope glass slide under a light microscope. These studies typically aim at detecting changes in cellularity or tissue architecture for the diagnosis of a disease. Over the last few decades, technological advances in scanning technology enabled the high-throughput conversion of glass slides into digital slides (whole-slide images) at resolutions approaching those of traditionally used optical microscopes. Digital pathology has become an active field that holds promise for the future of anatomic pathology and raises many pattern recognition research challenges such as rare object detection/counting and robust tissue segmentation [1,2,3]. In addition to the numerous potential patterns to recognize in digital slides, one of the key challenges for recognition algorithms is the wide variety of sample preparation protocols. These yield highly variable image appearances of tissue and cellular structures. Ideally, pattern recognition algorithms should be versatile so that they could be applied to several classification tasks and image acquisition conditions without the need to develop completely novel methods, but using training datasets related to each novel task at hand. However, such an idealistic application of pattern recognition methods on real-world applications requires the ground-truth data to be carefully designed and realistic. We believe realistic data collection is an underestimated challenge in digital pathology that deserves more attention. In this technical note, we first discuss potential dataset issues in digital pathology. We then suggest guidelines and tools to set up better ground-truth datasets and evaluation protocols.

Discussion

Potential Sources of Dataset Variability and Bias

Object recognition aims at designing methods to automatically find and identify objects in an image. The design of such methods usually requires ground-truth datasets provided by domain experts and depicting various categories (or classes) of objects to recognize. In object recognition research, publicly available ground-truth datasets are essential to enable continuous progress, as they also allow algorithm quantitative evaluation and comparison of algorithms. However, computer vision dataset issues have been raised recently against datasets used for several years [4-11]. We expect similar problematic issues might arise in the coming years in the emerging field of digital pathology if precautions are not taken when collecting new datasets.

Indeed, in some of these studies, published in the broader computer vision community, authors have shown that some hidden regularities can be exploited by learning algorithms to classify images with some success. For example, background environments can be exploited in several face recognition benchmarks [6,7]. Similarly, images of some object recognition datasets can be classified using background regions with accuracy far higher than mere chance [11] although images were acquired in controlled environments. In biomedical imaging, illumination, focus or staining settings might also discretly contribute to classification performance [10]. This type of fluctuation can lead to reduced generalization performance of classifiers as also observed in high-content screening experiments where images of different plates can have quite different gray value distributions [12].

Overall, these dataset biases will prevent an algorithm to work well on new images and are potentially guiding algorithm developers in the wrong direction. Moreover, the realism of several benchmarks has to be questioned beside the large amount of imaging data needed to analyze digital pathology applications. For example, in diagnostic cytology a single patient slide might contain hundreds of thousands of objects (cells and artifacts). However, typical benchmarks (e.g. [13] in serous cytology, and [14] in cervical cancer cytology screening) contain only a few hundred individual cells from a limited number (or unknown number) of patient samples, hence variations induced by laboratory practices and by biological factors are often not well represented. We believe that this partly explains why pattern recognition approaches had only a limited impact in cytology although there have been numerous attempts at designing computer-aided cytology systems [15].

The lack of details concerning data acquisition and evaluation protocols is also potentially hiding idiosyncrasies. An obvious sample selection bias would consist in collecting all examples of a given class (e.g. malignant cells) from a subset of slides while objects of another class (e.g. benign cells) are collected from another subset of slides. Such a data collection strategy might lead to classifiers that unwillingly capture slide-specific patterns rather than class-specific ones, hence have poor generalization performance. Similar problems might occur with other experimental factors, e.g. when examples from slides stained in a different laboratory or stained on different days of the week are used, as it has been shown that these are major factors causing color variations in histology [16]. It has been reported that many other factors (e.g. variation in fixation delay timings, changes in temperature, etc.) can affect cytological specimens [17] and tissue sections [18], hence the images used to develop recognition algorithms. Similarly, in immunohistochemistry, variable pre-analytical conditions (such as fluctuations in cold ischemia, fixation, or stabilization time) could induce changes on certain marker expression hence image analysis results [19]. Indeed, samples are prepared using colored histochemical stains that bind selectively to cellular components. Color variability is inherent to cytopathology and histopathology based on transmitted microscopy due to the several factors such as variable chemical coloring/reactivity from different manufacturers/batches of stains, coloring being dependent on staining procedures (timing, concentrations, etc.). Also, light transmission is a function of tissue section thickness and influenced by the components of the different scanners used to acquire whole-slide images.

Data Collection Guidelines

While it would be hardly possible to avoid all dataset variability and bias, it is important that the protocols for data acquisition and imaging acquisition try to reduce the non-relevant differences between object categories. Moreover, object recognition evaluation protocols should focus on challenging methods in terms of robustness.

Table 1 lists and organizes recommendations for less biased data collection based on lessons learned from the design of a practical cytology system [20], from observations in digital pathology challenges [21], from more general recommendations in the broader microscopy image analysis [22,23,35], and from computer vision literature [32]. While all these recommendations might not be followed simultaneously due to current standard practices and limited resources, we recommend to follow the most of these whenever possible.

Guideline category	Guideline description
Technical variabilities	Collect examples for each object category from different slide id / sample / day / technician / staining equipment / scanner and keep track of provenance to control hidden relationships. If the production environment is well controlled, include only those variations that will be encountered in the final application. Collect examples such that these variations are equally represented for each class. It includes acquiring images using the different slide scanners that will be used in the final application; on different days of the week and/or from different laboratories;
Biological variabilities	Cover variabilities (shape, texture, size, color, …) of objects within each category so that each category include a wide range of biological variations and not only examples corresponding to theoretical object's appearances. Also include an `others' category, as many non-cellular objects are often present in real-world samples (e.g. dust particles, bubbles, various contaminants) and might be found by automated object detection step. Classifiers not trained with negative examples might generate too many false positives;
Training set: class definition and sampling, object delineation	Match the object classes to the final application rather than to pathologist's textbooks. If the goal is to detect a specific type of rare cells, it might not be necessary to work on a multi-class definition of the task.

	Balance class distributions as much as possible and follow the experts' annotation process as they might annotate more `normal' objects (e.g. benign cells in cytology screening) due to their abundance. When class balancing is difficult (e.g. for rare cell detection tasks), consider data augmentation techniques afterwards; Instead of delineating objects of interest manually, consider the final whole-slide image analysis pipeline that will first apply a pre-processing steps (e.g. object detection using thresholding). Objects detected by this automated procedure should then be classified manually by experts to build the ground truth, so that training and testing sets are using the same kind of delineation procedure (rather than manual for training and automated in the final application);
Evaluation protocols and quality control	When reporting recognition performances, do not use cross-validation protocols that mix samples without taking into account their provenance. For example, cells from a single slide should not be both in training and test sets, otherwise robustness to new slides would not be properly assessed. Indeed, consider matching the final practical use of the system where experts analyze unseen slides. Objects coming from independant slides should therefore be kept out for validation. To evaluate methods, evaluation criteria adopted by the pattern recognition

	community can be used (classification accuracy, true positives, false negatives, F1-Score). However, accuracy evaluations might also be made on end-outcomes used by pathologists to better meet real-world expectations. Therefore, more task-specific statistical assessment might be adopted to tune hyper-parameters during learning and to test and compare recognition techniques. During the dataset creation, regularly control its quality (see next Section).
Reproducibility, traceability, and software tools	Provide fine details of the acquisition protocol when publishing a new dataset to allow reviewers to scrutinize it and identify potential sources of bias. Leverage existing open-source, collaborative, software and database to keep track of annotations performed by several experts and make clear accounts of the data collection methodology. To the best of our knowledge, Cytomine is the only open-source software that enable web-based and independent annotations of whole-slide images to collect and distribute large, semantic, ground-truth datasets [24];

Dataset Quality Control

While following guidelines for the construction of a realistic ground truth should reduce dataset bias, it might not be possible to control and constrain every aspect of the data collection due to current laboratory practices and available resources. Hence there might still be real-life reasons for dataset shift [29]. While other work have considered ground truth quality assessment using various annotation scoring functions (e.g. [30] where authors used the number of control points in the bounding polygon of a manual annotation), we believe these are not very relevant for practical pattern recognition applications in digital pathology. As [6,10], we rather think it is important to assess dataset quality with respect to end-outcomes used by final users. We therefore recommend to implement two simple quality control tests for assessing novel datasets, and detecting biases before intensively working on them.

The first strategy simply evaluates recognition performances (e.g. classification accuracy) with global color histogram methods or related approaches. While color information can be helpful for some classification tasks, too good results using such a simple scheme might reveal that individual pixel intensities are (strongly) related to image classes. In particular, in histology and cytology, color statistics may be of additional value e.g. to indirectly recognize a cell with a larger dark nucleus, but experts usually discriminate objects based on subtle morphological or textural criteria. For example, we have observed that staining variability can be exploited by such an approach on a dataset of 850 images of Hematoxylin and eosin (H&E) stained liver tissue sections from an aging study involving female mice on ad-libitum or caloric restriction diets [26]. We use the Extremely randomized Trees for Feature Learning (ET-FL) open-source classification algorithm of [25] that yields less than 5% error rate to discriminate mouse liver tissues at different development stages using only individual pixels encoded in the Hue-Salutation-Value (HSV) colorspace (using ten-fold cross-validation evaluation protocol, and method parameter values [25] were: T=10, nmin=5000, k=3, with NLs=1 million pixels extracted from training images). We observed that a similar approach yields also less than 5% error rate for the classification of 1057 patches of four immunostaining patterns (background, connective tissue, cytoplasmic staining, nuclear staining) from breast tissue microarrays [27] (using the same evaluation protocol and method parameter values).

Secondly, similarly to [6] that observed background artifacts in face datasets, one can easily evaluate recognition rates of classification methods on regions not centered on the objects of interest. We performed such an experiment using all 260 images of an acute lymphoblastic leukemia lymphoblasts (ALL) [28]. Using the ET-FL classifier [25], we obtained 9% error rate using only pixel data from a square patch of 50x50 pixels extracted at the top-left corner of each image corresponding to background regions (using ten-fold cross-validation evaluation protocol, and method parameter values [25] were: T=10, nmin=5000, k=28, with NLs=1 million 16x16 subwindows extracted from training images and described by HSV pixel values). That is significantly better than majority/random voting although these patches do not include any information about the cells to be recognized. This problem is illustrated in Fig 1.

In these two datasets, some acquisition factors are correlated to individual classes. Overall, these overly simple experiments stress the need for carefully designed datasets and evaluation protocols in digital pathology.

Conclusions

Pattern recognition could significantly shape digital pathology in the next few years as it has a large number of potential applications, but it requires the availability of representative ground-truth datasets. In this note, we summarized data collection challenges in this field and suggest guidelines and tools to improve the quality of ground-truth datasets. Overall, we hope these comments will complement other recent studies that provide guidelines for the design and application of pattern recognition methodologies [33,25,34], hence contribute to the successful application of pattern recognition in digital pathology.

Figure 1 Legend: Illustration of illumination/saturation bias in unprocessed images from a dataset describing normal and lymphoblast cells [28]. The large images (left) are two images from each class. Small images are cropped subimages (top left 50x50 corner) from 16 images for each class. Classifying these background subimages is much better than random guessing.

Reference:

1. Fuchs TJ, Buhmann JM. Computational pathology: Challenges and promises for tissue analysis. Journal of Computerized Medical Imaging and Graphics, 35(7):515–530, 2011.
2. Kothari S, Phan JH, Stokes TH, Wang MD. Pathology imaging informatics for quantitative analysis of whole-slide images. Journal of the American Medical Informatics Association, 20(6):1099–1108, November-December 2013.
3. McCann MT, Castro C, Ozolek JA, Parvin B, Kovacevic J. Automated histology analysis: opportunities for signal processing. IEEE Signal Processing, 2014.
4. Ponce J, Berg TL, Everingham M, Forsyth DA, Hebert M, Lazebnik S, Marszalek M, Schmid C, Russell BC, Torralba A, Williams CKI, Zhang J, Zisserman A. Toward

CategoryLevel Object Recognition, chapter Dataset Issues in Object Recognition. Springer-Verlag Lecture Notes in Computer Science, 2006.

5. Herve N, Boujemaa N. Image annotation: which approach for realistic databases? In Proc. ACM International Conference on Image and Video Retrieval (CIVR), pages 170–177, 2007.
6. Shamir L. Evaluation of face datasets as tools for assessing the performance of face recognition method. International Journal of Computer Vision, 79(3):225–230, 2008.
7. Kumar N, Berg AC, Belhumeur PN, Nayar SK. Attribute and Simile Classifiers for Face Verification. In IEEE International Conference on Computer Vision (ICCV), Oct 2009
8. Cox DD, Pinto N, Barhomi Y, DiCarlo JJ. Comparing state- of-the-art visual features on invariant object recognition tasks. In Proc. IEEE Workshop on Applications of Computer Vision (WACV), 2011.
9. Torralba A, Efros A. Unbiased look at dataset bias. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
10. Shamir L. Assessing the efficacy of low-level image content descriptors for computer-based fluorescence microscopy image analysis. Journal of Microscopy, 243(3):284–292, 2011.
11. Model I, Shamir L. Comparison of dataset bias in object recognition benchmarks. IEEE Access, 3(1):1953–1962, 2015.
12. Harder N, Batra R, Diessl N, Gogolin S, Eils R, Westermann F, König R, Rohr K. Large-scale tracking and classification for automatic analysis of cell migration and proliferation, and experimental optimization of high-throughput screens of neuroblastoma cells. Cytometry A. 2015 Jun;87(6):524-40.
13. Lezoray O, Elmoataz A, Cardot. A Color object recognition scheme: application to cellular sorting. Machine Vision and Applications 2003, 14:166–171
14. Marinakis Y, Dounias G, Jantzen J. Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. Computers in Biology and Medicine 2009, 39:69–78.
15. Bengtsson E, Malm P. Screening for cervical cancer using automated analysis for pap-smears. Comput Math Methods Med. 2014;2014:842037.
16. Bejnordi BE, Timofeeva N, Otte-Holler I, Karssemeijer N, van der Laak J. Quantitative analysis of stain variability in histology slides and an algorithm for standardization. Proc. SPIE - The International Society for Optical Engineering (9041), 2014.
17. Sahay K, Mehendiratta M, Rehani S, Kumra M, Sharma R, Kardam P. Cytological artifacts masquerading interpretation. Journal of Cytology, 30(4):241–246, 2013.

18. Mcinnes EF. Artefacts in histology. Comparative Clinical Pathology, 13(3):100–108, 2005.

19. Marien K. The search for a predictive tissue biomarker for response to colon cancer therapy with bevacizumab. Doctoral Thesis, Universiteit Antwerpen, 2016.

20. Delga A, Goffin F, Maree R, Lambert C, Delvenne P. Evaluation of cellsolutions bestprep(r) automated thin-layer liquid-based cytology papanicolaou slide preparation and bestcyte(r) cell sorter imaging system. Acta Cytologica, 58(5):469–77, 2014.

21. Giusti A, Claudiu D, Caccia C, Schmid-Huber J, Gambardella LM. A comparison of algorithms and humans for mitosis detection. In Proc. International Symposium on Biomedical Imaging (ISBI), 2014.

22. Kozubek M. Challenges and Benchmarks in Bioimage Analysis. Adv Anat Embryol Cell Biol. 2016;219:231-62.

23. Shamir L, Delaney J, Orlov N, Eckley DM, Goldberg IG. Pattern recognition software and techniques for biological image analysis. PLoS Computational Biology, 6(11), 2010.

24. Maree R, Rollus L, Stevens B, Hoyoux R, Louppe G, Vandaele R, Begon JM, Kainz P, Geurts P, Wehenkel L. Collaborative analysis of multi-gigapixel imaging data using cytomine. Bioinformatics (2016) 32 (9): 1395-1401.

25. Maree R, Geurts P, Wehenkel L. Towards generic image classification using tree-based learning: an extensive empirical study. Pattern Recognition Letters (2016), 74 (15): 17–23

26. Shamir L, Macura T, Orlov N, Eckely DM, Goldberg I G. Iicbu 2008 - a benchmark suite for biological imaging. In 3rd Workshop on Bio-Image Informatics: Biological Imaging, Computer Vision and Data Mining, 2008.

27. Niwas Swamidoss I, Kårsnäs A, Uhlmann V, Ponnusamy V, Kampf C, Simonsson M, Wählby C, Strand R. Automated classification of immunostaining patterns in breast tissue from the human protein atlas. J Pathol Inform. 2013; 4(Suppl): S14.

28. Donida Labati R, Piuri V, Scotti F. All-idb: the acute lymphoblastic leukemia image database for image processing. In IEEE International Conference on Image Processing (ICIP), September 2011.

29. Quionero-Candela J., Sugiyama M., Schwaighofer A., Lawrence N.D., Dataset Shift in Machine Learning. The MIT Press, 2008.

30. Vittayakorn S. and Hays J. Quality Assessment for Crowdsourced Object Annotations. In Proc. of the British Machine Vision Conference, pages 109.1-109.11. BMVA Press, September 2011.

31. A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (7) (1997) 1145– 935.

32. Krig S., Ground Truth Data, Content, Metrics, and Analysis. Chapter 7 in Book “Computer Vision Metrics”, Springer, 2016.

33. Janowczy A. and Madabhushi A., Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform. 2016; 7: 29.

34. Goodfellow I., Bengio Y., Courville, A., Chapter “Practical Methodology”, Deep Learning. MIT Press, 2016.

35. N. Jeanray,R. Maree, B. Pruvot, O. Stern, P. Geurts, L. Wehenkel, M. Muller. "Phenotype Classification of Zebrafish Embryos by Supervised Learning", PLoS ONE 10(1): e0116989, 2015.

Acknowledgements:

R.M. was supported by the CYTOMINE and HISTOWEB research grants of the Wallonia (DGO6 n°1017072 and n°1318185). R.M. is also grateful to persons Philipp Kainz (Medical University of Graz), Vannary Meas-Yedid (Pasteur Institute, France), and Nicky d’Haene (Hopital Erasme, Université libre de Bruxelles) for their comments on an earlier version of the manuscript.

Figure 1:

Figure 1 Legend: Illustration of illumination/saturation bias in unprocessed images from a dataset describing normal and lymphoblast cells [28]. The large images (left) are two images from each class. Small images are cropped subimages (top left 50x50 corner) from 16 images for each class. Classifying these background subimages is much better than random guessing.