Eurographics Workshop on Visual Computing for Biology and Medicine (2016)
H. Hauser, A. Hennemuth, and A. Lundervold (Editors)

PATHONE: From one thousand patients to one cell

A. Corvò[1] , M.A. Westenberg[1] , M.A. van Driel[2] and Jarke J.van Wijk[1]

[1]Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands
[2]Philips Research, Eindhoven, The Netherlands

Abstract

Digital Pathology is a recent clinical environment in which Electronic Health Records (EHRs), biopsy data and whole-slide-images (WSI) come together to provide pathologists the necessary information for making a diagnosis. Integration of this heterogeneous data into a single application is still one of the challenges in the evolution of pathology to a digital practice. While pathologists can perform diagnoses routinely on digital slides only, this is not the case in clinical research. For such purposes, the link between clinicopathological information of patients and images is essential. For example, image analysis researchers who develop automated diagnostic (support) algorithms need to select a representative set of slides to evaluate their methods. To achieve this, they need applications that combine cohort specification, slide image exploration, and selection of suitable images. We present the visualization tool PATHONE, which enables users to perform these steps on a single screen, integrating cohort and WSI selection.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications—

1. Introduction

Digital Pathology has been enabled by the introduction of whole slide imaging scanners, which provide high-resolution images of tissue slides. While pathologists have been reviewing such tissue slides through conventional light microscopes since the 17th century, the new technology aims to move their clinical practice and workflow to computers. There are still many challenges ahead to complete this transition.

One of these challenges is the integration of Electronic Health Records (EHRs), tissue slides, slide annotations, pathologist's notes, and pathology reports into a single application that supports both clinical practice and clinical research. Compared to radiology, the workstations used in pathology lack integration of these different sources of information, which generates delays and mistakes [Kru10]. Moreover, clinical researchers cannot easily identify a set of relevant cases (a cohort) to be used in a clinical study, because of these separate systems.

In this paper, we propose PATHONE, a tool that aims at the integration of PATHology data into ONE single application. Instead of working with the clinical information system to search for patients and then with slides manager software or Picture Archiving and Communicating Systems (PACS) for the matching images, users can perform all these operations via one interface. Our initial target users are (clinical) image analysis researchers, as they need to set up benchmark data sets to validate their methods [DCH*10, LHH*15, VKH*12]. The current approach is to formulate a request for slides to a technician/histologist in a pathology lab. An example of such a request is:

> I would like all *H&E* slides from patients matching the following profile: all female breast cancer patients, post-menopausal, age > 40, HER2-positive, M-0, Herceptin-treated.

In this request, the researcher defines a cohort, a group of subjects that share the same characteristics such as the ones listed. For these subjects, the corresponding tissue slides need to be retrieved. From this initial set of slides, the researcher selects a subset based on quality criteria, e.g., the slides should be properly stained, have no air bubbles, have no tears in the tissue, etc. The next step is to take the slides to the image analysis workstation, and run the algorithm and analyze the results. Generally, this workflow may take weeks, because each step is accomplished on a different system or in a different location, by different people. In PATHONE, the whole workflow can be performed efficiently in one system by the image analysis researchers themselves.

Our aim and contribution are the following:

Aim: Enable a clinical researcher to construct a cohort of relevant cases and retrieve the corresponding slides for image analysis on a single front-end application placed in a digital pathology environment.

Contribution: We look specifically at the cohort formulation problem linked to slide retrieval from the point of view of an image analysis researcher or domain expert. We demonstrate that PATHONE supports this research workflow on a dataset of one thousand breast cancer patients, collected from the Cancer Genome Atlas (TCGA) [TCGan]. Users can construct simple cohorts and navigate in only a few steps to a single tumor cell as detected by image analysis.

DOI: 10.2312/vcbm.20161278

2. Related work

In this section we briefly review visualization approaches for cohort selection, histology image exploration and content-based image retrieval. Cohort selection tools have been proposed in the visualization community to enhance and improve this process.

Applications as COQUITO [Jos16], INFUSE [KPB14], and CAVA [ZGP14] allow users to build a cohort in an iterative way, and they improve upon the query-based extraction of cohorts from a database. However, in clinical (image analysis) research, the link between cohort details and medical images is important. While these tools provide many options and good decision support, they do not provide access to medical images matching the cohorts, nor do they provide integration with image analysis tools.

Histology image exploration has been addressed in visualization research. GRAPHIE [DWHM15] is a recent visual analytics tool to explore, annotate, and discover relationships in histology image collections. It supports slide browsing driven by computed features. The relationships between the images are represented by a graph, where each node has a color representing the type of tissue in an image patch. Clusters become visible, which improve the annotation process and understanding of the whole collection. Unfortunately the tool does not scale up to large image collections. In a digital pathology environment where hundreds of tissue slides are scanned daily, this is a limiting factor regarding the efficiency of the workflow.

Zegami [Zeg16] is commercial software for high throughput visual image exploration. This solution focuses on image exploration and filtering by fields not related to clinical information. Hundreds of image thumbnails are shown, and the user can choose to lay them out in a grid, graph, table, or a map. Different filters can be applied but this solution does not take cohort construction into account.

Finally, there are content-based image retrieval systems (CBIR). These systems aim to combine low-level image processing technology with high-level semantic analysis of medical image content [THI03, ZWGB03]. Such systems are suitable to find slides similar to a given one, but do not support cohort construction either.

Visualization tools with the ability to combine cohort construction with whole-slide-image (WSI) exploration are still lacking. Specifically aimed at digital pathology, our tool aims to fill this gap.

3. Data and Tasks

In PATHONE, we use the breast cancer dataset from The Cancer Genome Atlas (TCGA) [TCGan]. This project started in 2005 collecting data concerning genomic mutations responsible for cancer and all the information that could be interesting for research studies, including tissue slides. We selected this dataset because of the quality and variety in attributes, and because image analysis on digital slides of breast tissue is a well-studied topic [VPvDV14]. Furthermore, most of the samples that are inspected in pathology labs are from breast cancer patients [BMP], and, therefore, this data set is of great interest to image analysis researchers.

The dataset consists of 1078 breast cancer patients of whom almost 2000 slides were collected. For each patient, there are 150 clinical categorical and numerical attributes. Multiple samples/biospecimen per patient were collected too (primary tumor/blood sample). There are over 2200 biospecimen, which are linked to the slides where they came from. This multiplicity leads to about 150 additional attributes overall. Tissue slides are available from the portal to enhance research studies on pathology data [GCS*]. We downloaded 150 slides whose sizes range from 200 MB to 2.7 GB. By convention, images captured with a resolution of .25 micron/pixel (mpp) are referred to as 40X, images captured with a resolution of .5 mpp are referred to as 20X, etc. A typical image is about 80,000 x 60,000 pixels, or 4.8Gp, and captured in 24-bit color, amounting to about 15GB of data before compression [PPP12, SCPP11]. The dataset covers the type of information that is necessary to make a diagnosis in a pathology lab or in a tumor board. It consists of:

Clinicopathological information. This concerns all information that characterizes patients, such as demographic data, comorbidities, pathology data as tumor stage and classification, and results from laboratory tests (cancer specific).

Tissue information. This concerns excisions and biopsies of tissue, which are called samples or biospecimen in histology. Each of them is divided into portions. From a portion, a glass slide is created.

Slide information. For a single patient, multiple glass slides can be generated. These are digitized and an identification code is assigned to each of them. Our dataset provides information about the tissue type on the slide. All slides are *H&E* stained, where H and E stand for the chemical compounds hematoxylin and eosin, respectively. This is one of the principal staining methods in histology, often the gold standard, which gives a characteristic pink color to the tissue and colors the cell nuclei blue (see Fig.1f for an example).

Radiology Images. Pathology and radiology form the core of cancer diagnosis [SAE*12]. In The Cancer Imaging Archive (TCIA) [CVS*13], radiology exam type information is available for about 139 TCGA patients that either went for breast MRI or for screening mammograms.

We interviewed three image analysis researchers with experience in pathology labs, to understand their way of working and their needs. From these interviews, we derived the following main tasks:

T1 Search a cohort that fits the research question in the clinical information system.
T2 Look for tissue slides in the PACS or the glass archive to find which blocks of tissue correspond to the relevant cases.
T3 Make a list of the tissue slides suitable for image analysis, based on their appearance and quality.
T4 Pre-process images and run the algorithms on digital slides. Example image analysis methods are nuclei detection, identification of tissue types, and detecting particular receptors through other kinds of staining. For further details concerning breast cancer image analysis, we refer to Vega et al. [VPvDV14].

The image analysis researchers indicated that the first three tasks suffer from multiple time-consuming checks in order to be sure that each slide actually matches the established criteria. A major problem is that these tasks have to be performed on different systems,

Figure 1: *The screenshot shows the main components of PATHONE. Each panel can be directly related to a task. **T1** can be performed on the selection panel (a). The matching cases and slides are visible in the query view (b) and the slides gallery (c) which correspond to **T2**. **T3** can be performed as the user moves to the options available in the slide filtering panel (d). Hence, **T2** and **T3** can be fulfilled in (c) and (a); views are updated according to the available images for the current selection. **T4** can be performed on the individual pop-up view shown in (e). Nuclei features are represented in the window (e) and results from detection become visible zooming on an interesting region (f)*

typically via some intermediate person like a lab assistant, which makes it cumbersome to refine search criteria efficiently.

4. Approach

Our approach consists of cohort selection and visualization, slide retrieval, individual image viewing, and a concept image analysis result visualization on a tissue slide of interest. The user interface consists of three sections; each one offers different functionalities and is designed for a different purpose. The screenshot shown in Fig.1 presents the main components of PATHONE. The *selection view* (a) is dedicated to select clinical (and pathological) features (**T1**), and provides a visualization of the selected cohort with respect to the whole data set. This component hosts four sections arranged vertically, which we explain in the following section. The *query view* (b) and the *slides gallery view* (c) show the current query and a thumbnail gallery of matching slides (**T2**). The *slide filtering view* (d) is dedicated to tissue slide filtering (**T3**). These three panels are linked views. Every action performed on *Panel a* affects our cohort selection and the gallery view. An example for **T4** is presented in a pop-up window accessible from the slides gallery. In the following, we explain the components of PATHONE in more detail.

4.1. Selection View

The *selection view* allows users to construct a cohort of patients and provides cohort visualization. This view contains several components to filter on attribute values:

Favorite Features. The user starts with **T1** from a preset of filters as shown in Fig. 1.a-1. Each attribute can have multiple values of which the semantics are clinically specific. The user has a list of preset values to use at first to build the cohort. According to the type of cancer, the researcher will need to set specific criteria. In Fig. 1, the preset list of attributes for breast cancer contains gender, tumor tissue site, pathologic stage, histological type, and attributes regarding the TNM staging system [LHS09].

A list of hundreds of attributes and values is difficult to explore, and, in this context, it becomes important to understand the effects of the current selection parameters. Therefore, for each attribute, the possible values are visualized in interactive rectangular boxes of which the inner fill level corresponds to the percentage of patients fulfilling the query parameters. Each box can act as a filter in the selection; a click on a box adds the attribute as a filter criterion, and affects the cohort selection. We chose for simple boxes as they are familiar to our target users. Alternatively, treemaps could have been used to obtain a more compact and detailed overview of the attributes (as in COQUITO [Jos16]) but they are less intuitive to use.

Features on demand. This panel (Fig. 1a-2) gives access to different categories of attributes: generic features, drug treatment, follow up details, new tumor-event cases, and cancer-specific features. Each category has its own attributes which can be used to further specify the research question during **T1**. This panel has two views: one shows the attributes, the other lists the possible values for the selected attribute. The list views are enriched with a bar that indicates the number of patients fulfilling the selection criteria. At-

tributes can be added to the preset filters with a right mouse button click for future use.

Other Sources. Researchers may need complementary data to select their cohort of patients. Examples are radiology images, previous pathology reports, patients which have been administered with a specific drug or went for a specific radiation therapy, to name a few. If available, these can be added as selection criteria (Fig. 1.a-3).

My Slides. A list of slides satisfying the filter criteria (**T2**) is shown in a separate tab (Fig. 1.a-4). Information about the tissue type is paired to the slide code. The user can click on the slide to see a larger version to assess the image's eligibility for the study (Not shown in Fig. 1).

4.2. Overview Component

The central area of PATHONE gives access to three different views dedicated to the output of users' actions.

My Query. Cohort specification (b) is shown in a table format. Each filter criterion (attribute name and value) corresponds to a row in this table. The patient and slide counts are shown as well. We decided to use a table to keep the query formulation clean and intuitive, while filters are applied. Our cohort selection does not support branching and joining in more complicated queries yet, hence a list view suffices. Filters can be deleted at any level. This action triggers an automatic update of the query according to the remaining filters, and it also updates the boxes of the attributes in the selection panel on the left. When a different value for the same attribute is selected the query is updated, and the visualization is updated accordingly. This component was designed to accomplish **T2**.

Slides Gallery. The slides are shown in a thumbnails grid. The user can navigate through the set of slides satisfying the criteria as desired in **T2**. Mousing over a thumbnail opens a tooltip containing more details of the image. Clicking an image opens a single-slide view in a pop-up window dedicated to the individual high-resolution digital slide.

Single-slide pop-up view. A slide viewer is commonplace in a digital pathology platform. We provide a viewer, which supports the standard zooming and panning features. In this view, the users can evaluate the quality of an image, inspecting for example the staining and the margins of the tissue (**T3**). Furthermore, this view provides a visualization of computed image features, if available. The figure shows an example of automated nuclei detection and a visualization of nuclei boundaries in yellow (**T4**).

4.3. Slide Filtering Component

Finding the right and most relevant tissue slides to construct a cohort is an important challenge (**T3**). In our dataset, each slide has information regarding the percentages of tissue type. This can be used, for example, to select images suitable for nuclei detection by filtering for slides with a high percentage of tumor cells and a low percentage of lymphocyte infiltration. The scatterplot gives visual support for this operation. In addition, the bar chart provides additional information about the slides, such as average percentage of tumor cells and tumor stages. The bars are interactive and can be clicked so that the corresponding attribute is added as a filtering criterion in the cohort selection. In future work, this component will be enriched with different tools which handle features generated from image analysis (e.g. object-level features [KPSW]), as shown in GRAPHIE.

4.4. Implementation

PATHONE is implemented in Java with a JavaFX user interface. To handle the tissue slides, we rely on the OpenSlide library [GGH*13]. This open source library is robust and shows high performance in handling and interacting with these large images.

5. Conclusion and Future work

In this paper, we have introduced PATHONE, a tool for cohort selection based on both clinical attributes and slide image attributes. We have identified a set of tasks specific to image analysis researchers in the context of digital pathology, and have shown that the corresponding research workflow can be performed with our tool. Although we have implemented our prototype on a dataset of breast cancer cases from TCGA, our approach is generic and directly supports other cancer datasets from TCGA, or other data sets with minor modifications.

The informal feedback collected from our three domain experts can be summarized as: "It can definitely speed up our research", "this would be really interesting for me","it would have been nice to have a similar tool in the last months". This shows that the current implementation of the tool satisfies their immediate needs. We plan to involve more domain experts, and perform a more formal evaluation.

As future work, we plan additional use cases coming from other domain experts (beyond image analysis). Furthermore, we will investigate statistical support and machine learning techniques to enhance the quality and the automation of the selection process. We also plan to incorporate more advanced (visual) cohort selection methods [Jos16] , [KPB14], [ZGP14] and [BLC*09]. An open problem is also the visualization of image analysis results given the size and scale of the image slides. Simple visualization by overlaying features and extracted edges leads to cluttered views, and specific solutions need to be developed to provide insight into these features across different resolutions.

6. Availability of the tool

Our tool is available at www.acorvo.it\pathone

References

[BLC*09] BUCUR A., LEEUWEN J. V., CHEN N.-Z., CLAERHOUT B., DE K., PEREZ-REY D., PARAISO-MEDINA S., ALONSO-CALVO R., MEHTA K., KRYKWINSKI C., NV C., POLITÉCNICA U., MADRID D., GROUP G. B.: Cohort Selection and Management Application Leveraging Standards-based Semantic Interoperability and a Groovy DSL. 4

[BMP] BRAY F., MCCARRON P., PARKIN D. M.: The changing global patterns of female breast cancer incidence and mortality. *Breast cancer research : BCR*, 6, 229–39. doi:10.1186/bcr932. 2

[CVS*13] CLARK K., VENDT B., SMITH K., FREYMANN J., KIRBY J., KOPPEL P., MOORE S., PHILLIPS S., MAFFITT D., PRINGLE M., TARBOX L., PRIOR F.: The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging 26*, 6 (dec 2013), 1045–57. doi:10.1007/s10278-013-9622-7. 2

[DCH*10] DOBSON L., CONWAY C., HANLEY A., JOHNSON A., COSTELLO S., O'GRADY A., CONNOLLY Y., MAGEE H., O'SHEA D., JEFFERS M., KAY E.: Image analysis as an adjunct to manual HER-2 immunohistochemical review: a diagnostic tool to standardize interpretation. *Histopathology 57*, 1 (jul 2010), 27–38. doi:10.1111/j.1365-2559.2010.03577.x. 1

[DWHM15] DING H., WANG C., HUANG K., MACHIRAJU R.: GRAPHIE: graph based histology image explorer. *BMC bioinformatics 16 Suppl 1*, Suppl 11 (jan 2015), S10. doi:10.1186/1471-2105-16-S11-S10. 2

[GCS*] GUTMAN D. A., COBB J., SOMANNA D., PARK Y., WANG F., KURC T., SALTZ J. H., BRAT D. J., COOPER L. A. D.: Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *Journal of the American Medical Informatics Association : JAMIA*, 6, 1091–8. doi:10.1136/amiajnl-2012-001469. 2

[GGH*13] GOODE A., GILBERT B., HARKES J., JUKIC D., SATYANARAYANAN M.: OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics 4*, 1 (jan 2013), 27. doi:10.4103/2153-3539.119005. 4

[Jos16] JOSUA KRAUSE, ADAM PERER H. S.: Supporting Iterative Cohort Construction with Visual Temporal Queries. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 91–100. 2, 3, 4

[KPB14] KRAUSE J., PERER A., BERTINI E.: INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data. *Visualization and Computer Graphics, IEEE Transactions on PP*, 99 (2014), 1. doi:10.1109/TVCG.2014.2346482. 2, 4

[KPSW] KOTHARI S., PHAN J. H., STOKES T. H., WANG M. D.: Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association : JAMIA*, 6, 1099–108. doi:10.1136/amiajnl-2012-001540. 4

[Kru10] KRUPINSKI E. A.: Optimizing the pathology workstation "cockpit": Challenges and solutions. *Journal of pathology informatics 1*, 1 (jan 2010), 19. doi:10.4103/2153-3539.70708. 1

[LHH*15] LAN C., HEINDL A., HUANG X., XI S., BANERJEE S., LIU J., YUAN Y.: Quantitative histology analysis of the ovarian tumour microenvironment. *Scientific reports 5* (jan 2015), 16317. doi:10.1038/srep16317. 1

[LHS09] LESLIE H. SOBIN MARY K. GOSPODAROWICZ C. W. (Ed.): *TNM Classification of Malignant Tumours, 7th Edition*. Wiley-Blackwell, 2009. 3

[PPP12] PARK S., PANTANOWITZ L., PARWANI A. V.: Digital imaging in pathology. *Clinics in laboratory medicine 32*, 4 (dec 2012), 557–84. doi:10.1016/j.cll.2012.07.006. 2

[SAE*12] SORACE J., ABERLE D. R., ELIMAM D., LAWVERE S., TAWFIK O., WALLACE W. D.: Integrating pathology and radiology disciplines: an emerging opportunity? *BMC medicine 10*, 1 (jan 2012), 100. doi:10.1186/1741-7015-10-100. 2

[SCPP11] SINGH R., CHUBB L., PANTANOWITZ L., PARWANI A.: Standardization in digital pathology: Supplement 145 of the DICOM standards. *Journal of pathology informatics 2* (jan 2011), 23. doi:10.4103/2153-3539.80719. 2

[TCGan] Last accessed on 2016 Jan. URL: https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp. 1, 2

[THI03] TANG H. L., HANKA R., IP H. H. S.: Histological image retrieval based on semantic content analysis. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society 7*, 1 (2003), 26–36. doi:10.1109/TITB.2003.808500. 2

[VKH*12] VETA M., KORNEGOOR R., HUISMAN A., VERSCHUUR-MAES A. H. J., VIERGEVER M. A., PLUIM J. P. W., VAN DIEST P. J.: Prognostic value of automatically extracted nuclear morphometric features in whole slide images of male breast cancer. *Modern Pathology* (2012), 1559–1565. doi:10.1038/modpathol.2012.126. 1

[VPvDV14] VETA M., PLUIM J. P. W., VAN DIEST P. J., VIERGEVER M. A.: Breast cancer histopathology image analysis: a review. *IEEE transactions on bio-medical engineering 61*, 5 (may 2014), 1400–11. doi:10.1109/TBME.2014.2303852. 2

[Zeg16] ZEGAMI: Zegami image explorer, 2016. URL: http://zegami.com/zegami-academics/. 2

[ZGP14] ZHANG Z., GOTZ D., PERER A.: Iterative cohort analysis and exploration. *Information Visualization* (2014), 1473871614526077. doi:10.1177/1473871614526077. 2, 4

[ZWGB03] ZHENG L., WETZEL A. W., GILBERTSON J., BECICH M. J.: Design and Analysis of a Content-Based Pathology Image Retrieval System. *IEEE Transactions on Information Technology in Biomedicine 7*, 4 (2003), 249–255. doi:10.1109/TITB.2003.822952. 2