Recent Advances in the Temple University Digital Pathology Corpus

I. Hunt, S. Husain, J. Simon[1], I. Obeid and J. Picone

The Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
{tuj85563, tuh26880, tug98850, iobeid, picone}@temple.edu

The Neural Engineering Data Consortium (NEDC) is developing a large open source database of high-resolution digital pathology images known as the Temple University Digital Pathology Corpus (TUDP) [1]. Our long-term goal is to release one million images. We expect to release the first 100,000 image corpus by December 2020. The data is being acquired at the Department of Pathology at Temple University Hospital (TUH) using a Leica Biosystems Aperio AT2 scanner [2] and consists entirely of clinical pathology images. More information about the data and the project can be found in Shawki et al. [3]. We currently have a National Science Foundation (NSF) planning grant [4] to explore how best the community can leverage this resource. One goal of this poster presentation is to stimulate community-wide discussions about this project and determine how this valuable resource can best meet the needs of the public.

The computing infrastructure required to support this database is extensive [5] and includes two HIPAA-secure computer networks, dual petabyte file servers, and Aperio's eSlide Manager (eSM) software [6]. We currently have digitized over 50,000 slides from 2,846 patients and 2,942 clinical cases. There is an average of 12.4 slides per patient and 10.5 slides per case with one report per case. The data is organized by tissue type as shown below:

Filenames:
tudp/v1.0.0/svs/gastro/000001/00123456/2015_03_05/0s15_12345/0s15_12345_0a001_00123456_lvl0001_s000.svs
tudp/v1.0.0/svs/gastro/000001/00123456/2015_03_05/0s15_12345/0s15_12345_00123456.docx

Explanation:
tudp: root directory of the corpus
v1.0.0: version number of the release
svs: the image data type
gastro: the type of tissue
000001: six-digit sequence number used to control directory complexity
00123456: 8-digit patient MRN
2015_03_05: the date the specimen was captured
0s15_12345: the clinical case name
0s15_12345_0a001_00123456_lvl0001_s000.svs: the actual image filename consisting of a repeat of the case name, a site code (e.g., 0a001), the type and depth of the cut (e.g., lvl0001) and a token number (e.g., s000)
0s15_12345_00123456.docx: the filename for the corresponding case report

We currently recognize fifteen tissue types in the first installment of the corpus. The raw image data is stored in Aperio's ".svs" format, which is a multi-layered compressed JPEG format [3,7]. Pathology reports containing a summary of how a pathologist interpreted the slide are also provided in a flat text file format. A more complete summary of the demographics of this pilot corpus will be presented at the conference.

Another goal of this poster presentation is to share our experiences with the larger community since many of these details have not been adequately documented in scientific publications. There are quite a few obstacles in collecting this data that have slowed down the process and need to be discussed publicly. Our backlog of slides dates back to 1997, meaning there are a lot that need to be sifted through and discarded for peeling or cracking. Additionally, during scanning a slide can get stuck, stalling a scan session for hours, resulting in a significant loss of productivity. Over the past two years, we have accumulated significant experience with how to scan a diverse inventory of slides using the Aperio AT2 high-volume scanner. We have been working closely with the vendor to resolve many problems associated with the use of this scanner for research purposes. This scanning project began in January of 2018 when the scanner was first installed. The scanning process was slow at first since there was a learning curve with how the scanner worked and how to obtain samples from the hospital. From its start date until May of 2019 ~20,000 slides we scanned. In the past 6 months from May to November we have tripled that number and how hold ~60,000 slides in our database. This dramatic increase in productivity was due to additional undergraduate staff members and an emphasis on efficient workflow.

The Aperio AT2 scans 400 slides a day, requiring at least eight hours of scan time. The efficiency of these scans can vary greatly. When our team first started, approximately 5% of slides failed the scanning process due to focal point errors. We have been able to reduce that to 1% through a variety of means: (1) best practices regarding daily and monthly recalibrations, (2) tweaking the software such as the tissue finder parameter settings, and (3) experience with how to clean and prep slides so they scan properly. Nevertheless, this is not a completely automated process, making it very difficult to reach our production targets. With a staff of three undergraduate workers spending a total of 30 hours per week, we find it difficult to scan more than 2,000 slides per week using a single scanner (400 slides per night x 5 nights per week). The main limitation in achieving this level of production is the lack of a completely automated scanning process, it takes a couple of hours to sort, clean and load slides. We have streamlined all other aspects of the workflow required to database the scanned slides so that there are no additional bottlenecks.

To bridge the gap between hospital operations and research, we are using Aperio's eSM software. Our goal is to provide pathologists access to high quality digital images of their patients' slides. eSM is a secure website that holds the images with their metadata labels, patient report, and path to where the image is located on our file server. Although eSM includes significant infrastructure to import slides into the database using barcodes, TUH does not currently support barcode use. Therefore, we manage the data using a mixture of Python scripts and manual import functions available in eSM. The database and associated tools are based on proprietary formats developed by Aperio, making this another important point of community-wide discussion on how best to disseminate such information.

Our near-term goal for the TUDP Corpus is to release 100,000 slides by December 2020. We hope to continue data collection over the next decade until we reach one million slides. We are creating two pilot corpora using the first 50,000 slides we have collected. The first corpus consists of 500 slides with a marker stain and another 500 without it. This set was designed to let people debug their basic deep learning processing flow on these high-resolution images. We discuss our preliminary experiments on this corpus and the challenges in processing these high-resolution images using deep learning in [3]. We are able to achieve a mean sensitivity of 99.0% for slides with pen marks, and 98.9% for slides without marks, using a multistage deep learning algorithm. While this dataset was very useful in initial debugging, we are in the midst of creating a new, more challenging pilot corpus using actual tissue samples annotated by experts. The task will be to detect ductal carcinoma (DCIS) or invasive breast cancer tissue. There will be approximately 1,000 images per class in this corpus. Based on the number of features annotated, we can train on a two class problem of DCIS or benign, or increase the difficulty by increasing the classes to include DCIS, benign, stroma, pink tissue, non-neoplastic etc.

Those interested in the corpus or in participating in community-wide discussions should join our listserv, *nedc_tuh_dpath@googlegroups.com*, to be kept informed of the latest developments in this project. You can learn more from our project website: *https://www.isip.piconepress.com/projects/nsf_dpath*.

Acknowledgments

This material is supported by the National Science Foundation under grants nos. CNS-1726188 and 1925494. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

This research would not be possible without the help of Dr. Nirag Jhala, MD, Professor of Pathology in the Lewis Katz School of Medicine and Laboratory Medicine Director, Anatomic Pathology/Cytology at Temple Hospital. He has been patiently teaching our undergraduates how to annotate pathology slides.

References

[1] D. Houser et al., "The Temple University Hospital Digital Pathology Corpus," in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 2018, pp. 1–7.

[2] Leica Biosystems, "Leica Biosystems Aperio AT2 – High Volume, Digital Whole Slide Scanning," *Leica Biosystems*, 2018. [Online]. Available: *https://www.leicabiosystems.com/digital-pathology/scan/aperio-at2/.*

[3] N. Shawki *et al.*, "The Temple University Digital Pathology Corpus," in *Machine Learning Application in Medicine and Biology (Tentative)*, 1st ed., I. Obied and J. Picone, Eds. New York City, New York, USA: Springer-Verlag, 2019, p. 45.

[4] I. Obeid and J. Picone, "CCRI: Planning: Digital Pathology Research Consortium." CISE Community Research Infrastructure (CCRI), National Science Foundation, January 1, 2020 – June 1, 2021.

[5] C. Campbell, N. Mecca, T. Duong, I. Obeid, and J. Picone, "Expanding an HPC Cluster to Support the Computational Demands of Digital Pathology," *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 2018, pp. 1–2.

[6] L. Biosystems, "Aperio ImageScope - Pathology Slide Viewing Software," *Leica Biosystems*, 2018. [Online]. Available: *https://www.leicabiosystems.com/digital-pathology/manage/aperio-imagescope/.*

[7] "Aperio Format," *OpenSlide*, 2018. [Online]. Available: https://openslide.org/formats/aperio/.

NEURAL ENGINEERING
DATA CONSORTIUM
www.nedcdata.org

Recent Advances in the Temple University Digital Pathology Corpus

I. Hunt, S. Husain, J. Simons, I. Obeid and J. Picone

The Neural Engineering Data Consortium, Temple University

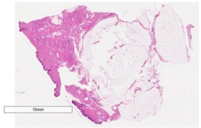
College of Engineering
Temple University

Abstract

- The Neural Engineering Data Consortium (NEDC) is developing a large open source database of high-resolution digital pathology images known as the Temple University Digital Pathology Corpus (TUDP) .
- The data is being acquired at the Department of Pathology at Temple University Hospital (TUH) using a Leica Biosystems Aperio AT2 scanner.
- The data consists entirely of clinical pathology images collected over the last decade at TUH.
- Our team scans approximately 2,000 slides a week and has scanned a total of ~60,000 slides.
- The average image size is 100 MB. The total size of the corpus is currently 6 Terabytes. The final 1M corpus is expected to be 0.1 Petabytes.
- In a pilot study, we achieved a mean sensitivity of 99.0% for automatic identification of slides with pen marks, and 98.9% for slides without marks.
- We are currently developing a second pilot corpus of slides containing 3,600 examples of breast cancer studies. We expect about 30% to have evidence of cancer. Annotations consist of four labels: ductal carcinoma (DCIS), invasive carcinoma (INVC), neoplastic (NNEO) and normal (NORM).
- Our long-term goal is to release a database of one million high-resolution digital pathology slide images over the next decade; we expect to release the first 100,000 images by December 2020.

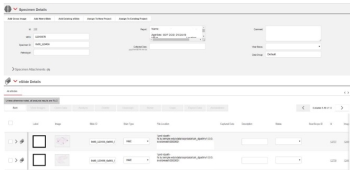
About the Scanner

- The Aperio AT2 requires at least eight hours of scan time to scan a full load of 400 slides.
- With three undergraduate workers (30 hours per week of labor), we scan about 2,000 slides a week.
- The efficiency can vary greatly depending on the quality and condition.
- We have reduced the slide fail rate from approximately 5% to 1% by optimizing how we handle the slides.



About sSlide Manager (eSM)

- Aperio's workflow product to manage a database of pathology slides for clinical applications



Specimen Details

Add Gross Image | Add New eSlide | Add Existing eSlide | Assign To New Project | Assign To Existing Project

MRN: 12345678
Specimen ID: 0s00_123456
Report: Name:
Age/Sex: 66/F DOB: 01/24/49
Data Group: Default

Specimen Attachments (1)

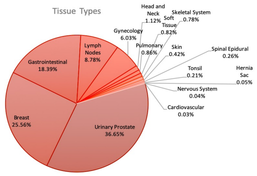
eSlide Details

Unless otherwise noted, all analysis results are RUO

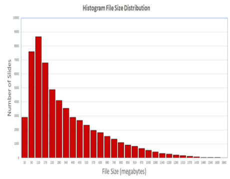
Label	Image	Slide ID	Stain Type	File Location	Captured Date	Description	Status	ScanScope ID	Id
		0s00_123456_0a000_l	H&E	\\pnd-dpath-fs.tu.temple.edu\data\isipdata\tuh_dpath\v1.0.0\svs\breast\000000\					12277
		0s00_123456_0a001_l	H&E	\\pnd-dpath-fs.tu.temple.edu\data\isipdata\tuh_dpath\v1.0.0\svs\breast\000000\					12278

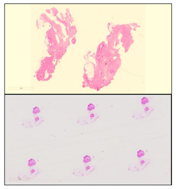
- Every specimen has a patient report attached as an MS Word document. The raw text from the report is searchable from within eSM.

About the Data



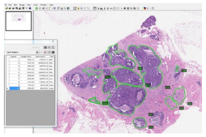
- The AT2 scanner automatically saves images in an .svs file format.
- .svs files are TIFF files in which images of various resolutions have been layered and compressed with lossy compression.
- Base image resolution is high: 50K x 50K pixels.
- Data is organized by tissue type. We currently identify 15 tissue types.
- The average digitized slide requires about 100 MB of space, which translates to 10M images per Petabyte.
- Each slide can have multiple images, which makes the machine learning problem much harder.

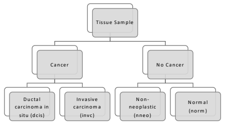




Annotations and Corpora

- ImageScope is our primary tool for annotation.
- It provides tools for adjusting and inspecting the annotations. The pen tool is most used to indicate and label one of the four classes, and a negative pen tool complements the pen tool to allow for certain areas to be ignored (not included in the annotation).
- Our next release will be a pilot corpus of 3,600+ breast cancer slides because the diagnosis of this is a challenging task for humans.
- This dataset will be a four-class problem:



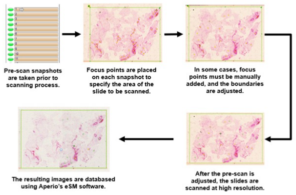


- Each slide will typically contain 5 to 10 regions that were influential in making the diagnosis. An image is viewed and analyzed at a rate of 5 minutes per slide, allowing for ~50 annotations in a 4-hour period.
- After an image has been annotated, an XML file is generated which contains the annotation data. For the annotation to appear on the image, the XML file must be present with the same image name and in the same folder as the original .svs image.

Optimizing Our Workflow While Maintaining HIPAA Compliance

Scanning Slides:

- About 30% of our older slides (more than 10 years old) are set aside due to peeling or cracking.
- Slides are cleaned and loaded into the scanner.
- Snapshots, which are lower resolution preview images, are taken to optimize the scanning process.
- We crop these images, so we only scan the exact region of interest.



Pre-scan snapshots are taken prior to scanning process.

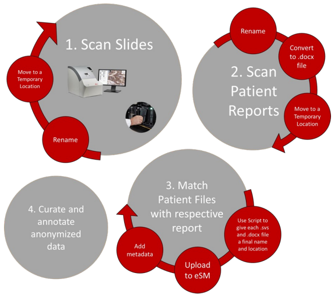
Focus points are placed on each snapshot to specify the area of the slide to be scanned.

In some cases, focus points must be manually added, and the boundaries are adjusted.

After the pre-scan is adjusted, the slides are scanned at high resolution.

The resulting images are databased using Aperio's eSM software.

- To help reduce the slide fail rate, a blank slide is scanned to correct the contrast ratio.
- Focus points are manually adjusted, which is time-consuming, to improve the quality of the scans.



1. Scan Slides
Move to a Temporary Location
Rename

2. Scan Patient Reports
Rename
Convert to .docx file
Move to a Temporary Location

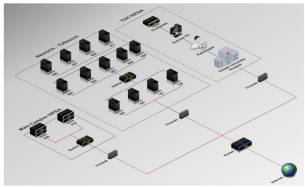
3. Match Patient Files with respective report
Use Script to give each .svs and .docx file a final name and location
Upload to eSM
Add metadata

4. Curate and annotate anonymized data

Curation and Anonymization of the Data:

- Due to data interchange issues, we scan hardcopies of the reports, convert them to MS Word, and manually correct scanning errors while deidentifying the report.
- We manually pair the reports with the slide images and upload the report into eSM.
- Patient metadata, which includes the stain type, filename, location and patient MRN, is also added to eSM.
- The computer network that supports this project spans three secure networks. The eSM database reside on Temple Hospital's HIPAA network.

Databasing and Archival:



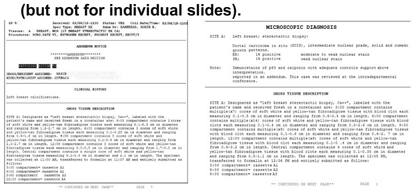
Renaming and Storage Conventions

- Data is organized so it can be easily accessed using standard Unix commands:

```
/data/tudp/v1.0.0/svs/gastro/001234/00123456/
2015_03_05/0s19_12345/0s19_12345_0a001_12345678_lvl0001_s000.svs
```

	Directory Components:		
Field	**Description**	**Template**	**Example**
database name	4-letter acronym	NNNN	tudp
version	v<major>.<minor>.<sub>	vx.x.x	v1.0.0
file type	type of data	fff	svs
case	6-letter code	cccccc	gastro
sequential ID	6-digit directory ID	######	001234
patient ID	8-digit patient ID	########	00123456
date	date specimen was collected	yyyy_mm_dd	2015_03_05
specimen type	4-digit code	tt##_	0s19_
sequence number	5-digit sequence number	#####	12345
	Filename Components		
specimen type	4-digit code	tttt_	0s19_
sequence number	5-digit sequence number	#####	12345
block level ID	2-letter code + 3-digit sequence	ll###	0a001
patient ID	8-digit (zero-padded) patient ID	########	00123456
block cut type	3-letter code + 4-digit sequence	bbb####	lvl0001
sequence number	3-digit sequence	s###	s000
file extension	three-letter filename extension	.ext	.svs

- Anonymized pathology reports are available in a Word and flat text file for each sequence number (but not for individual slides).



```
SP #: [redacted]        Received: 03/08/18-1230  Status: VER   Coll Date/Time: 03/08/18-1100
                        Spec Type: BREAST BX      Subm Dr: BANKHEAD, ROBIN R.
Tissues:  A   BREAST, NOS (LT BREAST STEREOTACTIC BX CA)
Procedures: SURG.PATH VI, ESTROGEN RECEPT, PROGEST RECEPT, RECUT/5

                          ADDENDUM NOTICE

                     ********ADDENDUM********
                       SEE ADDENDUM DATA SECTION

     PATHOLOGY ASSIGN TO

GROSS/RESIDENT ASSIGNED:  NAZIR [redacted]
MICRO/PATHOLOGIST ASSIGNED: JITURAJA [redacted]

                          CLINICAL HISTORY

Left breast calcifications.

                      GROSS TISSUE DESCRIPTION

SITE A: Designated as "left breast stereotactic biopsy, Ca++", labeled with the
patient's name and received fresh in a coretainer are: 3:00 compartment contains 2 cores
of soft white and yellow-tan fibroadipose tissue each measuring 0.1-0.3 cm in diameter
and ranging from 1.2-2.7 cm in length. 6:00 compartment contains 3 cores of soft white
and yellow-tan fibroadipose tissue each measuring 0.1-0.25 cm in diameter and ranging
from 0.4-1.6 cm in length. 9:00 compartment contains 3 cores of soft white and
yellow-tan fibroadipose tissue each measuring 0.1-0.5 cm in diameter and ranging from
2.0-2.7 cm in length. 12:00 compartment contains 3 cores of soft white and yellow-tan
fibroadipose tissue each measuring 0.1-0.5 cm in diameter and ranging from 1.7-3.0 cm in
length. Central compartment contains a single core of soft white and yellow-tan
fibroadipose tissue measuring 0.1-0.3 cm in diameter and 2.1 cm in length. The specimen
was collected at 11:00 AM, transferred to formalin at 11:57 AM and entirely submitted as
follows:
3:00 compartment= cassette A1
6:00 compartment= cassette A2
9:00 compartment   cassette A3
12:00 compartment= cassette A4

         ** CONTINUED ON NEXT  PAGE**            PAGE   1
```

```
                      MICROSCOPIC DIAGNOSIS

SITE A:   Left breast; stereotactic biopsy:

          Ductal carcinoma in situ (DCIS), intermediate nuclear grade, solid and comedo
          growth patterns.
          ER:       1% positive        moderate to weak nuclear stain
          PR:       1% positive        weak nuclear stain

Note:     Immunostains of p63 and calponin with adequate controls support above
          interpretation.
          reported in an addendum. This case was reviewed at the intradepartmental
          conference.

                      GROSS TISSUE DESCRIPTION

SITE A: Designated as "left breast stereotactic biopsy, Ca++", labeled with the
patient's name and received fresh in a coretainer are: 3:00 compartment contains
multiple(x7) cores of soft white and yellow-tan fibroadipose tissue with blood clot each
measuring 0.1-0.5 cm in diameter and ranging from 0.6-3.4 cm in length. 6:00 compartment
contains multiple(x14) cores of soft white and yellow-tan fibroadipose tissue with blood
clots each measuring 0.1-0.4 cm in diameter and ranging from 0.5-2.6 cm in length. 9:00
compartment contains multiple(x9) cores of soft white and yellow-tan fibroadipose tissue
with blood clot each measuring 0.1-0.6 cm in diameter and ranging from 0.6-2. 7 cm in
length. 12:00 compartment contains multiple(x9) cores of soft white and yellow-tan
fibroadipose tissue with blood clot each measuring 0.1-0 .4 cm in diameter and ranging
from 0.4-3.2 cm in length. Central compartment contains 4 cores of soft white and
yellow-tan fibroadipose tissue with blood clots each measuring 0.1-0.35 cm in diameter
and ranging from 0.5-1.1 cm in length. The specimen was collected at 12:05 PM,
transferred to formalin at 12:36 PM and entirely submitted as follows:
3:00 compartment= cassette A1
6:00 compartment= cassette A2
9:00 compartment= cassette A3

         ** CONTINUED ON NEXT  PAGE**            PAGE   1
```

Conclusion

- Based on the current rate of production, we expect to scan 80,000 slides in the year 2020 and release 100,000 images by December 2020.
- We maintain a mutually beneficial relationship with TUH. Pathologists provide us with slides, and we manage the AT2 scanner and eSM to aid their work. They are now using digital slides in their tumor board reviews and are increasingly incorporating digital imaging into their workflow.
- In 2020, we are executing an NSF planning grant titled "CCRI: Planning: Digital Pathology Research Consortium." The goal of this grant is to explore community-wide resources that will enable machine learning applications in digital pathology. To learn more about this project, email help@nedcdata.org.

Acknowledgements

- This material is supported by the National Science Foundation under grants nos. CNS-1726188 and 1925494. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
- This research would not be possible without the help of Dr. Nirag Jhala, MD, Professor of Pathology in the Lewis Katz School of Medicine and Laboratory Medicine Director, Anatomic Pathology/Cytology at Temple Hospital. He has been patiently teaching our undergraduates how to annotate pathology slides.