Cervical Cancer Detection with Synthetic Data and CNNs

Sean Wade

A project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

David Wingate, Chair
TODO
TODO

Department of Computer Science

Brigham Young University

ABSTRACT

Cervical Cancer Detection with Synthetic Data and CNNs

Sean Wade
Department of Computer Science, BYU
Master of Science

Cervical cancer is one of the deadliest cancers amoung women worldwide. There are approxiamtly 270,000 deaths a year, out of which 85% occure in developing countries. The goal of this project is to help in the development of low cost sensors and new algorithms to detect cervical cancer in India. Early diagnosis of this cancer is extreamly effective in treatment.

Recent advances in neural networks can be applied to the early detection of cervical cancer. The primary bottleneck is the lack of labeled data and cost of generating it. This project will focus on new ways of generating this data and using it in state of the art computer vision algorithms.

Keywords: segmentation, CNN, synthetic data, data augmentation, cervical cancer, mediacl imaging

Table of Contents

Chapter 1

Introduction

Cervical cancer accounts for 6.6% of cancer cases in the world. In 2018 alone, there were well over 570,000 cases. Even worse is the disproportionate way populations are affected. About 85% of deaths from cervical cancer came from lot to middle income countries.[1]

Fortunatly the high mortality rate from cervical cancer can be reduced by effective screening and early treatment.

1.1 The pap-smear screening

The pap-smear screening was developed by Georges Papanicolaou. Using a small brush, a cytological sample is taken from the cervix and smeared onto a thin glass slide. To clarify the cells characteristics, the smear is stained using the Papanicolaou method. This emphisizes the different components of the cells with specific colors, making it more clear in a microscope.[4] Each microscope slide contains up to 300,000 single cells with defferent orientations and overlap[4]. This has made automatic segmentation methods challenging.

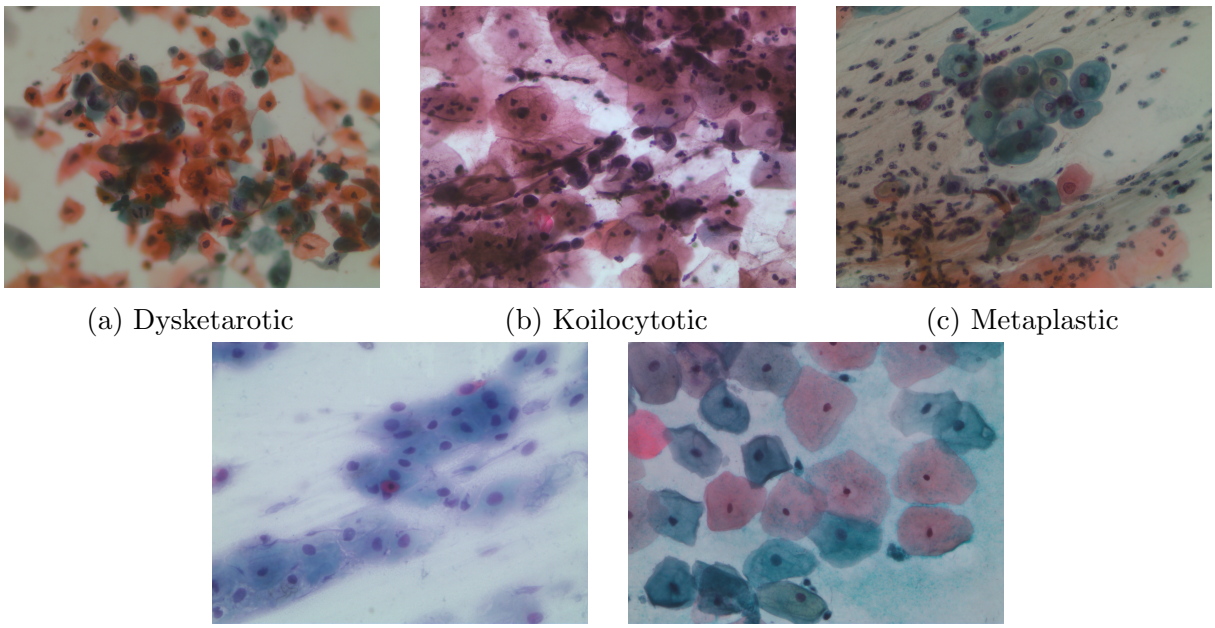
Chapter 2

Datasets

The datasets used are constructed by cyto-technicians for classification purposes. These technicians use a microsope witha resolution of 0.201 µm / pixel to grab every cell.

2.1 SIPaKMeD

The SIPaKMeD dataset consists of 996 cluser cell images of Pap smear slides. From these there are 4049 indavidual cells that are segmented. These cells are then labeled in 5 categories: (a) Dyskeratotic, (b) Koilocytotic, (c) Metaplastic, (d) Parabasal and (e) Superficial-Intermediate.



(a) Dysketarotic (b) Koilocytotic (c) Metaplastic

(d) Parabasal (e) Superficial-Intermediate

Figure 2.1: Cell Types

Superficial-Intermediate cells constitute the majority of the cells found in a Pap test. Usually they are flat with round, oval or polygonal shape cytoplasm stains mostly eosinophilic or cyanophilic. They contain a central pycnotic nucleus. They have well defined, large polygonal cytoplasm and easily recognized nuclear limits (small pycnotic in the superficial and vesicular nuclei in intermediate cells). These type of cells show the characteristics morphological changes (koilocytic atypia) due to more severe lessions.

Parabasal cells are immature squamous cells and they are the smallest epithelial cells seen on a typical vaginal smear. The cytoplasm is generally cyanophilic and they usually contain a large vesicular nucleus. It must be noted that parabasal cells have similar morphological characteristic with the cells identified as metaplastic cells and it is difficult to be distinguished from them.

Koilocytotic cells correspond most commonly in mature squamous cells (intermediate and superficial) and some times in metaplastic type koilocytotic cells. They appear most often cyanophilic, very lightly stained and they are characterized by a large perinuclear cavity. The periphery of the cytoplasm is very dense stained. The nuclei of koilocytes are usually enlarged, eccentrically located, hyperchromatic and exhibit irregularity of the nuclear membrane contour.

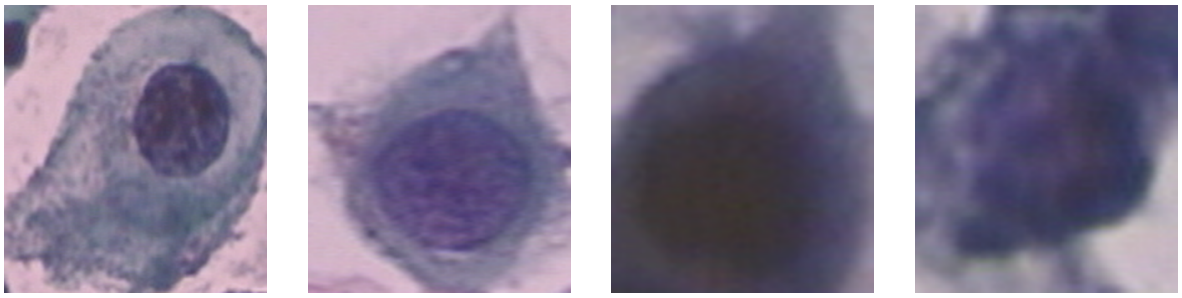
Dysketarotic cells are squamous cells which undergone premature abnormal keratinization within individual cells or more often in three-dimensional clusters. They exhibit a brilliant orangeophilic cytoplasm. They are characterized by the presence of vesicular nuclei, identical to the nuclei of koilcytotic cells. In many cases there are binucleated and/or multinucleated cells.

Metaplastic Cells are small or large parabasal-type cells with prominent cellular borders, often exhibiting eccentric nuclei and sometimes containing a large intracellular vacuole. The staining in the center portion is usually light brown and it often differs from that in the marginal portion. Also, there is essentially a darker-stained cytoplasm and they exhibit great

uniformity of size and shape compared to the parabasal cells, as their characteristic is the well defined, almost round shape of cytoplasm.[5]

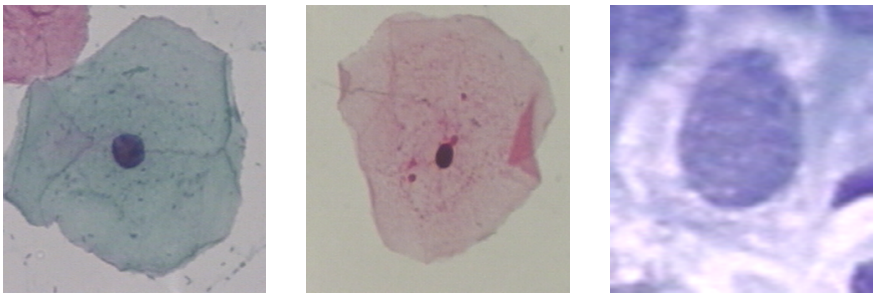
2.2 Herlev

The Herlev dataset is comprised of 917 isolated single cell images. These are distributed unequally between seven classes of cells. Superficial squamous epithelia, intermediate squamous epithelia, columnar epithelial, mild squamous non-keratinizing dysplasia, moderate squamous non-keratinizing dysplasia, severe squamous non-keratinizing dysplasia and squamous cell carcinoma in situ intermediate.[2]



Light Dysplastic | Moderate Dysplastic | Severe Dysplastic | Carcinoma in Situ

Figure 2.2: Abnormal Cells



Intermediate | Superficiel | Columnar

Figure 2.3: Normal Cells

Chapter 3

Background

Current automatic screening scystems that incorperate pap smear have a very similar workflow: cell segmentation, cytoplasm and nuclei segmentation, feature extraction, and cell classification.

Whole slide classification (no segmentation) [3]

Chapter 4

Synthetic Data Generation

- rotations
- flip horizontal and vertical
- random crop and zoom

Chapter 5

Segmentation Model

Chapter 6

Experiments and Reuslts

Chapter 7

Future Work

Chapter 8

Tools and skills learned

Project Requirments

8.1 Project Initiation

The student finds a mentor who agrees to work with him/her on a project not later than the end of his/her 3rd semester in the program. The project may be initiated by the student or the mentor.

8.2 Project Development and Completion

When the student is ready to complete the project, the student signs up for CS 698R (add code should be obtained from the Graduate Academic Advisor). The student will be expected to put 125-150 hours of work into the project during that semester. CS 698R is a pass/fail class. It is expected that students will only take this class once. The proposal form must be completed, signed and turned into the Graduate Academic Advisor by the first day of classes in the semester/term you intend to take the course.

The student prepares a short proposal and submits it to the graduate coordinator by the end of the first week of the semester. The proposal focuses on specific activities and clearly-defined deliverables, approved by the mentor and then by the graduate coordinator. The proposal should include a checklist of deliverables that will be demonstrated in the final presentation. The mentor and student meet at least weekly throughout the project so that the student has a meaningful mentoring experience. The mentor must approve all revisions to the project deliverables checklist. The student tracks hours worked on the project and keeps a week-by-week log of hours spent working on the project, split into categories such as design, coding, writing, etc. This 1) allows the mentor to make sure that expectations are met, 2) helps curb project creep, and 3) is common practice in industry (particularly if working for a

company that handles multiple consulting/contracts/projects, where it is important to track to the fraction of an hour how long is spent on each). By the middle of the semester (by the end of the 9th week), the student produces a draft report, also focused on deliverables, a copy of which must be submitted to the Graduate Coordinator. An email will be sent to all faculty mentors and MS Project students by the end of 7th week of the semester as a reminder of this requirement. The student and mentor discuss the draft report, review the time log, go over the write-up, and make scope adjustments if needed. This contributes to the student having a meaningful and culminating writing experience. Additional reviewing iterations are recommended, but optional. The student prepares a final written scientific project report, including at least an executive summary, an introduction, a project description, a validation section, and a conclusion. At the end of the semester, project presentations are held within a 3-hour block during finals week. Two members of the Graduate Committee, together with the mentor for each project, make up the student's committee. The Committee sits at a desk with all the project reports in a binder in order of student presentation. The students present for 15 minutes after which there are 5 minutes for questions and comments. The Committee reviews the time log, and checks off deliverables as the student presents. If all deliverables are satisfactory, the student passes. If not, the Committee may discuss and advise as to whether the student passes or fails. Project Report The MS project report document should be submitted to the committee at the end of the semester in which the student takes CS 698R. The document must be about 15 double-spaced pages. The project report should provide necessary background and then argue that a significant piece of work was needed. The contributions should reflect the importance of the work. A summary of the project's time log must also be included.

8.3 Project Report Presentation

Oral Presentation Audience: CS faculty members who may not be acquainted with the topic.

A 12-15 minute oral presentation of the project must be carefully organized and given to the members of the MS committee and the invited public. During the project report presentation, the student must answer committee member's questions on such areas as method, significance, organization, and literature search. After the presentation, the student and public leave the room while the committee comes to a decision on project report.

Examination Results:

At this point the examining committee decides on a result. The possible results are:

Pass

Pass with qualifications - Revision to project is an example of why this would be selected.

Fail - Fail the oral exam and be terminated from the graduate program.

The final CS 698R grade will be determined by the Advisor

References

[1] World health organization. cervical cancer, 2018. [online; accessed 12-december-2018].

[2] Jan Jantzen, Jonas Norup, Georgios Dounias, and Beth Bjerregaard. Pap-smear benchmark data for pattern classification. In *Proc. NiSIS 2005*, pages 1–9. NiSIS, 2005.

[3] Kranthi Kiran GV and G Meghana Reddy. Automatic classification of whole slide pap smear images using cnn with pca based feature interpretation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[4] Jonas Norup. Classification of pap-smear data by tranduction neuro-fuzzy methods. 2005.

[5] Marina Plissiti, P. Dimitrakopoulos, Giorgos Sfikas, Christophoros Nikou, O. Krikoni, and Antonia Charchanti. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. pages 3144–3148, 10 2018. doi: 10.1109/ICIP.2018.8451588.