CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Raghav Ganesh

Project Number

S0809

Project Title

A Novel, Fast, and Accurate Machine Learning Based Cell Classification Algorithm for Pathology Labs to Diagnose Diseases

Abstract

Objectives/Goals

Physicians often rely on histology & cytology lab reports to diagnose diseases. Lab report generation needs manual intervention which adds cost and time. These prevent a local point of service in remote, rural, and underdeveloped areas.

Objective: Use Machine Learning and Computer Vision to develop a cell segmentation algorithm (to detect and discriminate individual cells from a microscope slide image), a robust cell classification algorithm (to classify the individual cell images), and demonstrate these on images captured off of microscope slide using a frugal add-on to a cell phone camera.

Design constraints: Algorithm over 80% accuracy validated on over 10,000 cell images, test time under 30 sec/cell, and the image capturing attachment cost under $10.

Methods/Materials

I developed my machine learning algorithm (coded in Java, C++, Batch, Python using OpenCV, SciPy and CImg libraries) using an iterative approach (4 revisions, each using results from prior versions). CUDA integration speeded up the training and segmentation process. I captured images of cheek cell swabs from volunteers using my $3 DIY macro lens cell phone camera attachment.

Training: 1st version - Identified the pros and cons of 9 classification algorithms (K Nearest Neighbors, Linear SVM, RBF SVM, Gaussian, Decision Tree, Random Forest, Neural Net, Adaboost, and QDA). 2nd version - Added iterative thresholding before feeding the data to the classifiers. 3rd version - Focused on detecting FAST & SIFT, and feature matched using FLANN and a FAST Knn classifier, assigned each feature a weight. 4th version - Generated feature descriptors from the locations of FAST and SIFT features, grouped similar features together using agglomerative clustering, and trained Knn classifiers for each cluster.

Testing: Grayscaled the images, performed coarse edge detection, and tested the watershed method, blob detection, k means clustering for cell segmentation. Each cluster#s classifier was weighted and run.

Results

Achieved a classification accuracy of 91% (with 10K cell images) and test time of 1.65 sec/cell (60% saved with CUDA). Data was divided randomly into 10 groups, each run with 90% train-10% test split.

Summary Statement

Developed a novel machine learning algorithm to classify cells (10,000 images) with an accuracy of 91% and processing time of 1.65 sec/cell (60% reduction with CUDA) and tested on cheek-cell images captured using my macro-lens phone add-on.

Help Received

I am thankful to the contributors to the various online forums and classes, these helped me ramp up on Machine Learning. I thank Prof. Lezoray (U of Caen) for letting me use his images database, and science teacher Mr. Lee.