Parallel Massive Dataset Cleaning

Jianmeng Yu

Supervised by: Bob Fisher

Introduction and Motivation

- What is the project about?
  - Translate the existing decision algorithms into Python.
  - Develop a framework to apply the algorithm at a massive parallel scale.
  - Running the cleaning algorithm and reconstruct the video files.
- Why parallel?
  - The original code would take about 25,000 hours on a 40-core machine to finish.
  - There are about 839,000,000 frames in 396,000 video clips.
  - And 1,446,000,000 records in a 500 GB .sql file.

Current Progress - Code Translation

One of the original goal of the project is to translate the Pipeline Classifier developed by Matthew Pugh, from Matlab to Python. This Pipeline classifier is still at experiment stage, it's not assembled together, but all of the component is implemented.

- What's the translation progress?
  - Most of the component is rewritten with Python.
  - Feature Extraction in Matlab was called using PyMatlab library.
  - Neural Network written in Lua Torch was called using Lutorpy library.
- Why using a bridge on these part instead of translating?
  - Some part of the code were customized and was not documented.
  - Estimated time cost of translating the code is larger than the time it could reduce.

Current Progress - Applying Cleaning Algorithm

The task distribution is implemented using MPI (message passing interface). Some minor fixes were made to the framework created last semester, to prevent thrashing on DICE machines.
Due to usage of the lab during day, the cleaning algorithm is only running during night, on about 200 DICE machines in Appleton Tower.
It took about 20 days for the most costly feature extraction part to finish. The only remaining part of the pipeline is classification using SVM and CNN, which would take about a week to finish.

Future Work - Finishing Cleaning Algorithm

After finishing the classification, the only part left would be combining the results to generate a decision vector of each video. Since the Pipeline Classifier is only at experiment stage, the final combine stage used is not decided. Some evaluation on the classification results may be needed to produce the final result. Also, the training data used is heavily biased, if the outcome of the algorithm have a poor accuracy on unseen dataset, re-training might be needed.
If the classification is producing sensible results, then the videos will be reconstructed to a smaller set.

Future Work - Scheduling

- What's the plan in week 5-11?
  - Week 5-6: Finish the remaining cleaning, write coursework and dissertation while waiting the code to finish.
  - Week 7-11: Evaluate the cleanness of the algorithm, Finish the dissertation. Reconstruct the videos.
- What if the algorithm perform poorly on unseen data?
  - Create new Ground Truth Dataset.
  - Experiment on different classifiers.