# Parallel Massive Dataset Cleaning

Jianmeng Yu

Supervised by: Bob Fisher

# Background - Dataset

- Fish4Knowledge (F4K)[3] collected 5 years of fish video.
- They reduce the recording to 1.6 TB size.
- 500 GB `.sql` dump, 1.1 TB video and summary csv files.

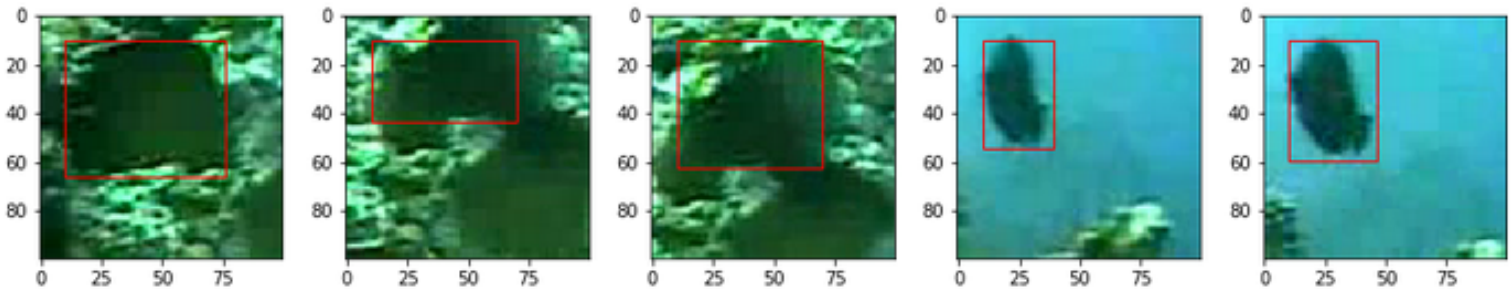


Figure: Sample frames of the videos

# Background - Project Motivation

- 1.6 TB dataset, 839 million detections.
- About 60% of estimated False Positives.
- Pugh[1]'s prototype cleaning algorithm.
- Not applied due to cost (800,000 hours)[2].

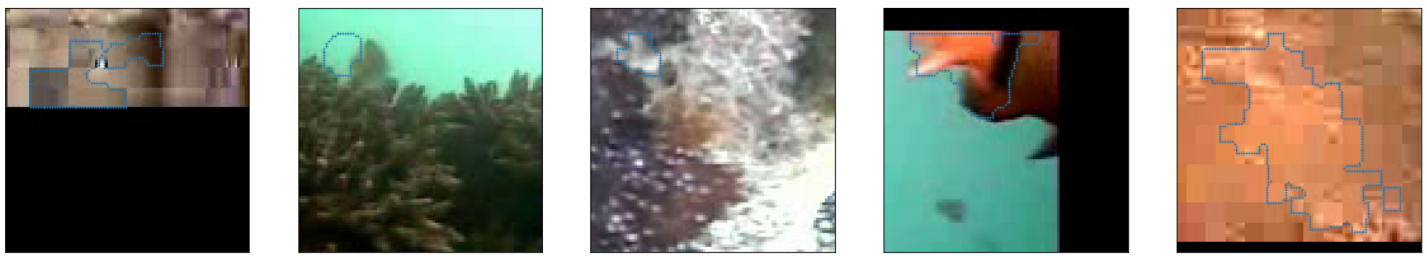


Figure: False Positives in the Dataset

# Background - Project Goals

- Apply the cleaning algorithm to remove False Positive.
- Need a parallel framework to reduce time.
- Need to translate the code to Python.

Figure: True Positives in the Dataset

# Background - Cleaning Algorithm

- Developed by Matthew Pugh[1].
- Use Python script to extract info from video/summary.
- FEIF to remove some partial-fish before classify.
- Feature extraction, transformation for SVM.
- Transformation and normalize image for CNN.
- Use Top-N algorithm to obtain decision vector.

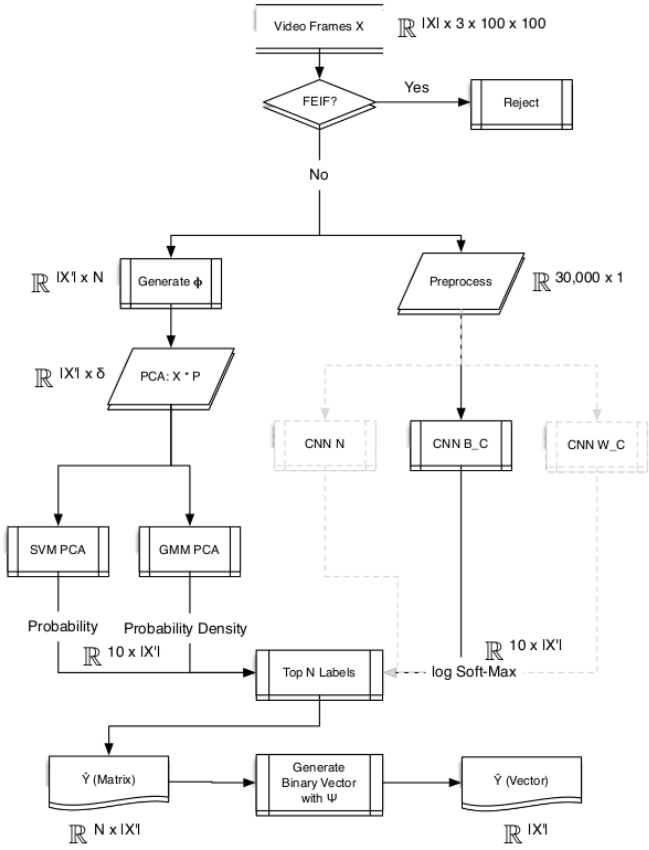



Figure: Pipeline Classifier

# Project Problems

The project has encountered several problems with this algorithm:

- The pipeline is only tested without parallelization.
- Dump file top large for school PostgreSQL service.
- Ambiguous classification schema.
- Training data wasn't representative.
- Extreme runtime cost - even for 200 DICE machines.
- Translation needed, for more portability.
- CNN trained is broken, over-fits on badly transformed dataset.
- Prototype stage - The best Top-N strategy was not found.

# Details - Frame Edge Indicator Function

FEIF removes all of fish touching frame edge.

- Reduces dataset by 10%.
- Fast, but reject good fish sometimes.

Another similar method is used to reject planktons.

- Reject all the videos recorded at night.
- These videos have over 99% of False Positive.
- Reduces dataset by 8%.

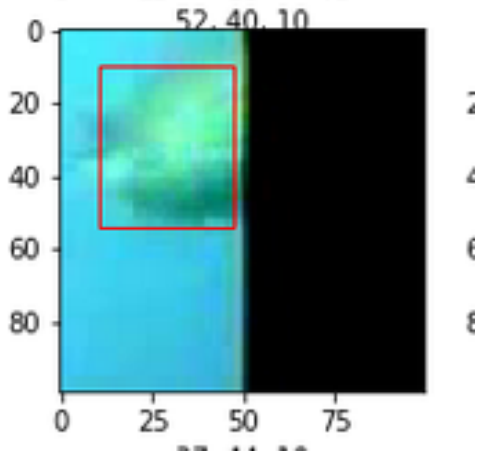



Figure: Sample Frame Touching Edge

# Details - Ground Truth

- Pugh created a 10-class classification schema.
- He marked 60,000 detections for training.
- Class 6 and 8 are acceptable fishes.

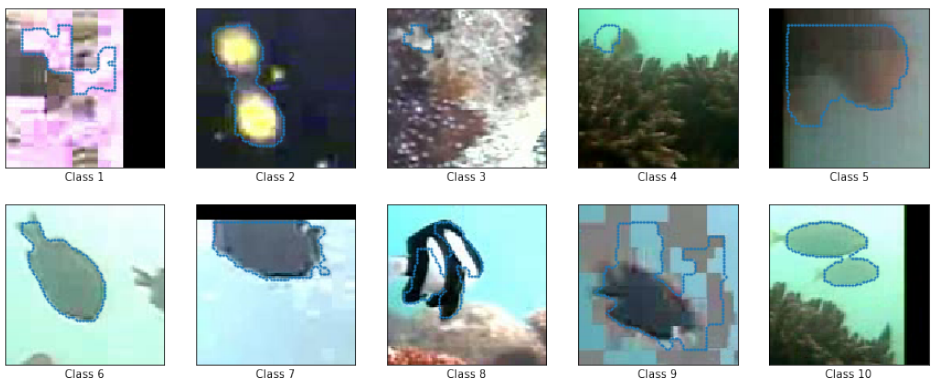



Figure: Sample fish from each class

- This project marked another 20,000 detections for validation.

# Details - Classifiers

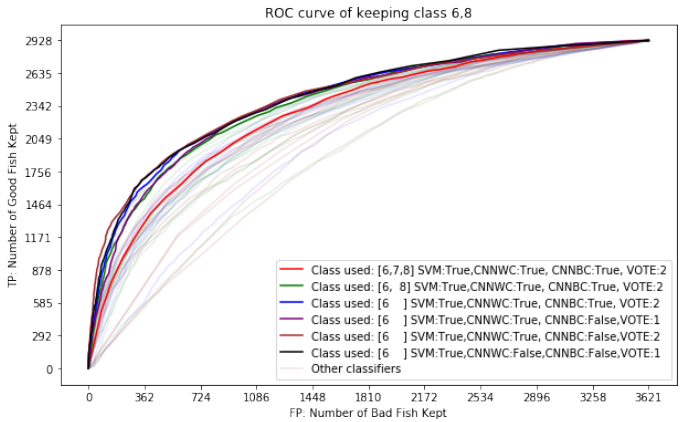
- The SVM performed as expected.
- CNNs training does not work as intended.
- This is caused by color space issues in OpenCV.
- Retraining the CNN too costly and not used.

Figure: Normal Image in RGB space

Figure: OpenCV output, in BGR space

# Details - Voting Strategy

- Uses voting instead of the original Top-N decision.
- Use ROC curve to pick strategy.
- For keeping most True Positive, use SVM only is the best strategy.



# Details - Translating MATLAB code into Python

- Most of the pipeline parts are translated into Python.
- Except Feature Extraction in MATLAB and CNN in Lua.
- After benchmarking, a full translation does not seem optimal.
- PyMatlab and Lutorpy are used instead.

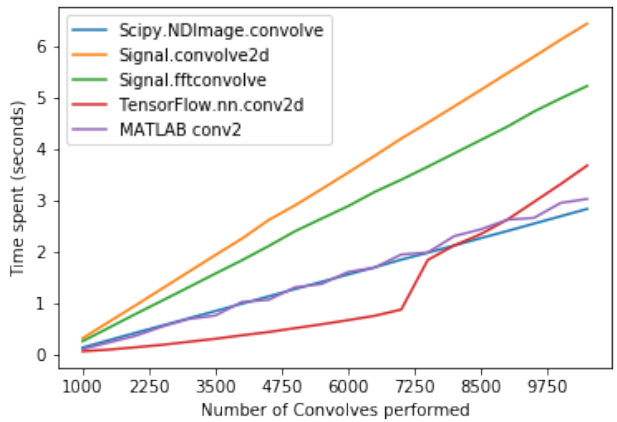



Figure: Performance of different convolution algorithms

# Details - Standard I/O Stream Extraction

Another problem is the lack of a SQL server.

- 500 GB `.sql` dump file too large for PostgreSQL service.
- Independent records, SQL query may not be best choice.
- Alternatively, parse relevant information to 400,000 csv files.
- Saves time and cost for maintaining a server.
- Reduces dump file to 2/3 size, also makes loading faster.

# Details - Parallel Distribution

- 220 student lab DICE machines is used.
- 880 cores reduces runtime significantly.
- A Python script is used to “scan” for idle machines.
- University provides a shared file system (AFS).
- MPI is used for the communication between process.

```
[binda]s1413557: mpiexec -n 4 -host binda,bertoglio,berzin,adorni python ~/Desktop/ms/example1.py
I am  adorni.inf.ed.ac.uk rank 3 (total 4)
I am  berzin.inf.ed.ac.uk rank 2 (total 4)
I am  bertoglio.inf.ed.ac.uk rank 1 (total 4)
I am  binda.inf.ed.ac.uk rank 0 (total 4)
  Slave berzin.inf.ed.ac.uk rank 2 executing "Do task" task_id "1"
  Slave bertoglio.inf.ed.ac.uk rank 1 executing "Do task" task_id "0"
Master: slave finished is task and says "I completed my task (0)"
  Slave adorni.inf.ed.ac.uk rank 3 executing "Do task" task_id "2"
Master: slave finished is task and says "I completed my task (1)"
```

Figure: Framework Example Run

# Parallel framework's time reduction

Each detection took about 1 second per core to process:

- Reduced from 8 second per core after translation.
- Dataset have 839,000,000 detections.
- This would take 2400 days on 4-core machine.
- Parallel Distribution reduces this to 12 days.

# Result - Outcome

Project Result:

- 28% of the 1.6 TB dataset is deleted.
- 40% of the False Positive rejected.
- 10% of the True Positive lost in the process.
- A binary decision array marking keep/reject for each video.

Does not achieve Pugh's predicted 90% FP removal:

- CNN over-fit, accuracy drops 20% on validation set.
- Poor coverage on training-set.
- Over-estimation due to Pugh's 2-Fold Training/Testing split.

# Result - Successful Samples

- The SVM appears to learn the shape of detections well.
- High probability detections are usually good fish.


Figure: Positives with High Probability

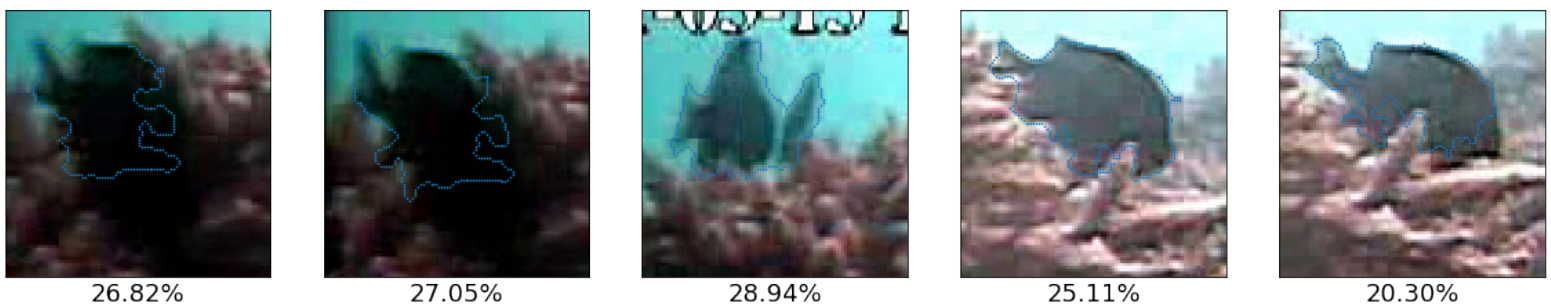



Figure: Positives with Low Probability

# Result - Failure Samples

- FP are mixed with TP with erratic contour.
- Fishes may have very erratic boundaries.
- Noises may have fish-like boundaries.

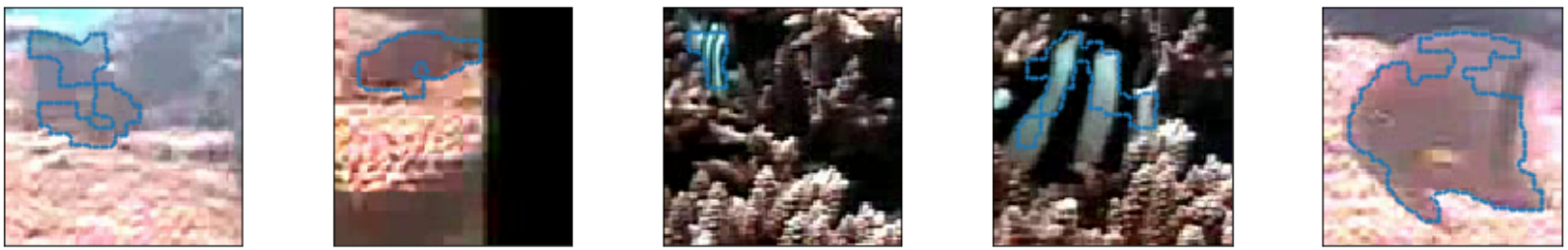


Figure: Good Fish Rejected due to Erratic Boundary (less than 1%)


Figure: False Positives kept with Very Low Probability

# Result - Achievements

- Finishes the 800,000 hour task in 100,000 CPU hours.
- Potentially reduces dataset by 500 GB.
- Removes 40% of FP at cost of 10% TP.
- Most kept FP have low SVM score.
- Most reject TP have very erratic boundary.

# Future Works

- Larger Ground Truth dataset: Better training.
- Fix the True Negatives caused by FEIF.
- Re-train the CNN used with new dataset.

# Questions?

# Bibliography

[1] Matthew Pugh.
Removing false detections from a large fish image data-set.
Msc dissertation, The University of Edinburgh, 2015.

[2] Qiqi Yu.
Adding temporal constraints to a large data cleaning problem.
Msc dissertation, The University of Edinburgh, 2016.

[3] Robert B Fisher, Yun-Heh Chen-Burger, Daniela Giordano, Lynda Hardman, Fang-Pang Lin.
*Fish4Knowledge: collecting and analyzing massive coral reef fish video data*.
Springer, 2016.