CS6240 FINAL PROJECT

Ramkishan Panthena

Karthik Subramanian

Venkatesh Koka

DECIDING TRAIN/VALIDATION SPLITS

	Background	Foreground	Baseline
Image 1	955179	10202	98.9432152
Image 2	976979	5292	99.4612485
Image 3	971664	6489	99.3366068
Image 4	972048	6296	99.3564636
Image 6	964472	8288	99.1479913

PRE-PROCESSING

- Dense Representation
  - 3087 features
- Sparse Representation
  - Keep only pixel with values > 20
- Rotations on foreground data
- Reducing class imbalance by sampling background records (1:1, 5:1, 20:1)

EXPERIMENTS

- Random Forests (Dense Representation)
  - TreeCount - 5
  - MaxDepth - 30
  - MaxBins - 128
  - Accuracy - 99.7204 (Baseline - 99.4612485)

EXPERIMENTS (CONTD.)

- Random Forests (Sparse Representation)
  - TreeCount - 100
  - MaxDepth - 30
  - Accuracy - 99.7634 (Baseline - 99.4612485)

EXPERIMENTS (CONTD.)

- Random Forests (Sparse Representation)
  - SubSamplingRate – 0.3
  - TreeCount - 100
  - MaxDepth - 30
  - Accuracy - 99.7232 (Baseline - 99.4612485)

EXPERIMENTS (CONTD.)

- Random Forests (Sparse Representation)
  - Rotations
  - TreeCount - 100
  - MaxDepth - 30
  - Accuracy - 99.5210 (Baseline - 99.4612485)

EXPERIMENTS (CONTD.)

- Random Forests (Sparse Representation)
  - TreeCount - 15
  - MaxDepth - 30
  - Sampled background records - 1:1, 5:1, 20:1
  - Accuracy - 98.8734, 99.5365, 99.7288 (Baseline - 99.4612485)

EXPERIMENTS (CONTD.)

- Gradient Boosted Trees (Sparse Representation)
  - Iterations - 100
  - MaxDepth - 2
  - Accuracy - 99.7383 (Baseline - 99.4612485)

EXPERIMENTS (CONTD.)

- Naïve Bayes (Sparse Representation)
  - Accuracy - 93.5234 (Baseline - 99.4612485, didn't beat baseline!)
  - Predicted most true positives out of all models but also predicted lot of false negatives

FINAL RUN

- Split the data into 3 subsets
- Trained a separate model on each partition:
  - Random Forest: 100 trees, 30 depth, sparse representation
  - Boosted Trees: 100 iterations, 2 depth, sparse representation
  - Boosted Trees: 50 iterations, 2 depth, 20:1 background-foreground split, sparse representation
- Use majority vote to decide the final prediction

RESULTS

- Model 1 : 99.7634
- Model 2 : 99.7383
- Model 3 : 99.7335
- Cumulative: 99.7405

FUTURE IMPROVEMENTS?

- Perform more experiments by reducing the class imbalance
- Train model by providing more weightage to less dominant class
- Logistic Regression, SVM with kernels (spark ml-lib limitations!)

THANK YOU