Progress Report 2: Image Classification Corn Kernels

Progress Thus Far:

Due to limited time this week I was not able to make many substantial steps in the process, though I was able to begin seeing results with a KNN model as well as doing some early tuning of the KNN model.

In order to better grasp the nature of the HOG (histogram of oriented gradients feature extraction method I looked into a means of visualizing the resulting features extracted. The results are essentially small lines with different orientations evenly spaced throughout the image. These lines possess the key features of the images in terms of the topography of an object within an image. Below in figures 1 and 2 there are 4 kernels visualized in greyscale as well as above their HOG visualization. Being able to have some visual representation of the feature extraction method seemed valuable to me in order to better grasp what information it emphasized.

Beyond my study of HOG I produced some predictions using a basic KNN model with number of neighbors set to 10. I used the feature matrix produced by the HOG method from sklearn as the training data in my data split, with the labels provided in the data set as the target variable. These early results can be seen in the below table:

	Precision	**Recall**	**F1-Score**	**Support**
Broken	0.7500	0.0749	0.1362	921
Discolored	0.1429	0.0020	0.0039	507
Pure	0.4168	0.9914	0.5868	1159
Silkcut	0.5556	0.0180	0.0348	278
Accuracy	-	-	0.4272	2865

Upon inspecting the resulting predictions I found that the majority of distributions simply belonged to the majority class, which explains the rather high recall seen for the pure class. The predictions per class can be seen in figure 3 below.

In order to test for the best possible value for number of neighbors I used a five fold cross validation split with shuffled indices. I used the range of neighbors: [1, 3, 5, 7, 9, 11, 13, 15, 17, 20, 25, 35, 50] of which the best accuracy score was ~.44 at 3 neighbors. This was consistent after multiple tests. Interestingly there was very little variance in the prediction accuracy at different numbers of neighbors, as the lowest accuracy score was ~.41 at 50 neighbors.

The fact that predictions were very heavily weighted towards the class with the highest count of examples in the data–meaning the data is not balanced–implied that further preprocessing would be necessary to improve the model. At this point I unfortunately have not been able to pursue these advancements any further but do have plans on several different strategies to attempt in coming weeks.

Next Steps:

There are several steps which I intend to attempt, each of which may be done in different combinations in order to produce the best possible results. First I will be doing feature reduction on the feature matrix generated via HOG with PCA in order to see if noise elimination may help improve results, this step will take some experimentation in terms of the number of components to retain upon training the model. Beyond feature reduction I think it may be reasonable to manually balance the data set in order to improve performance of the KNN model. By reducing noise from imbalanced data and PCA together the nearest neighbor approach may become more meaningful, as points in hypothetical n-dimensional space would hopefully begin to meaningfully cluster, instead of being somewhat randomly distributed.

After testing basic feature reduction I would like to focus on tuning the hyperparameters of HOG, there are several of which I would like to experiment with. The hyperparameters for the number of orientations as well as the pixels per cell and cells per block are all tunable parameters which I will test against for performance improvement. I believe the PCA step of preprocessing is likely to yield better results, though I would like to experiment at each stage of the preprocessing and feature extraction aspects of this project in order to build better intuition about the tools I am using.

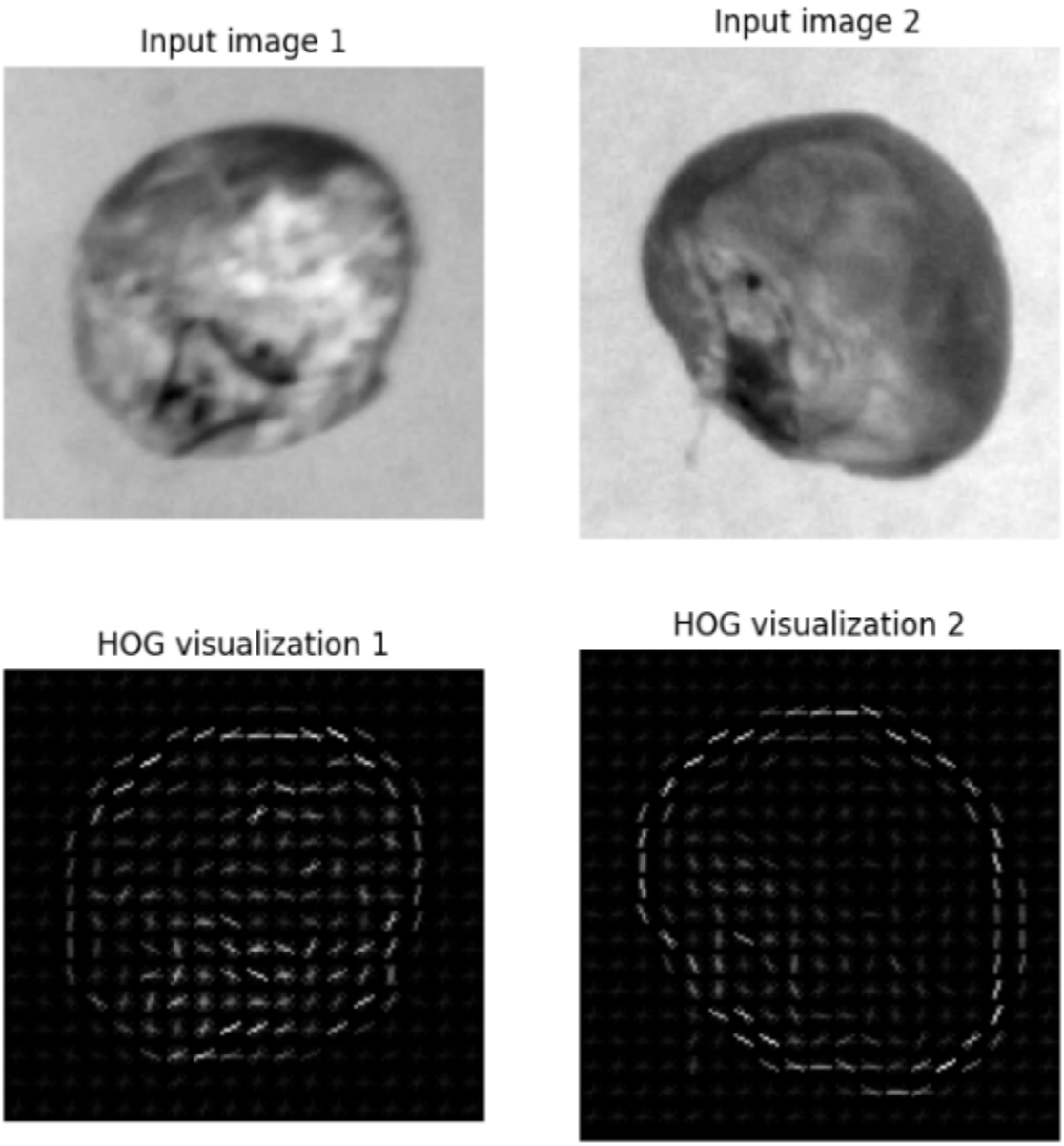
There are other feature extraction techniques which I intend to learn about and test as well, as mentioned previously SURF and SIFT may be better suited to the task I am attempting. As it stands I understand that HOG was initially developed for object detection and specifically human detection which suggests that it may not be best suited for picking out the subtleties of differences in corn kernels. Thus I believe that exploring alternative feature extraction methods is crucial in order to identify which methods have the correct strengths for the multi-class classification task of such similar objects.

Beyond alternative feature extraction methods, after I feel I have reasonably exhausted my options for improving a simple KNN model I will move onto different models as mentioned previously. Both the random forests model as well as the support vector machine model are options of interest to me at this point, and if I have enough time I would like to explore beyond them as well.

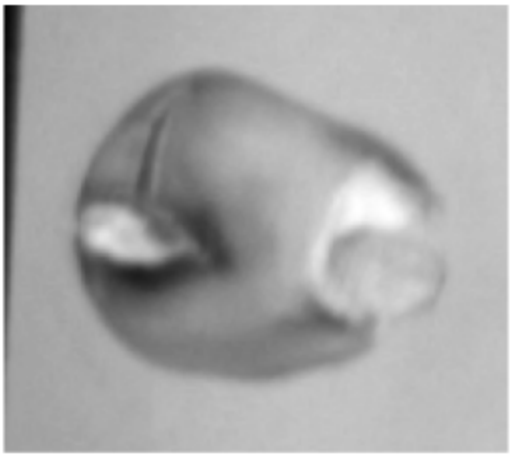
A core interest in this project for me is taking the time to learn the methods and tools I use intuitively such that I retain long term impressions about their strengths, weaknesses, and key differences. Which is why I am not emphasizing great results at the start, as I believe it would be more constructive for me to take the time necessary to understand these tools and their nuances at this early introductory phase.

Figures:

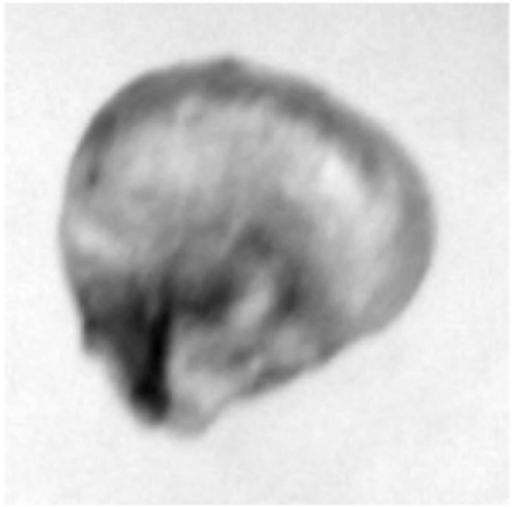
Figure 1-2:



Input image 3



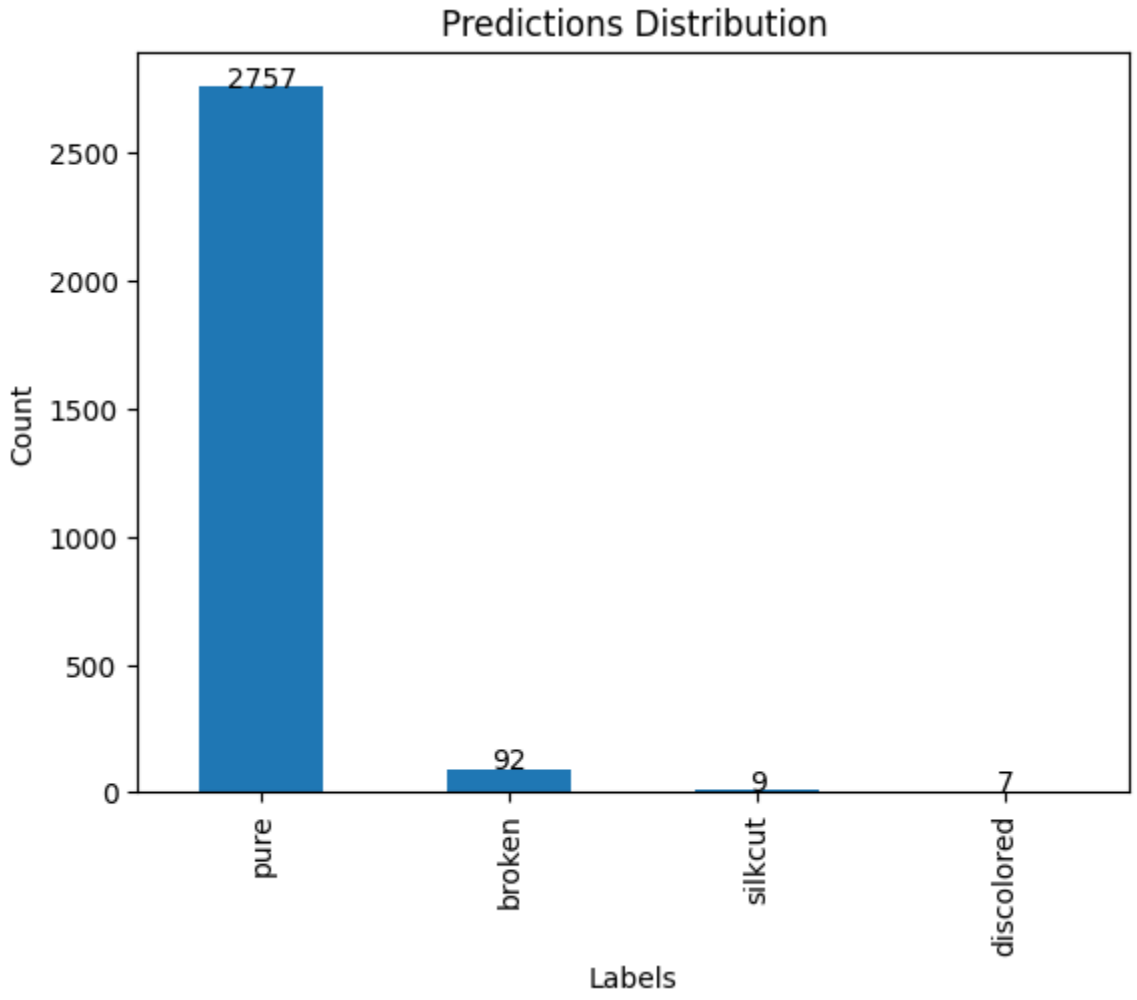
Input image 4



HOG visualization 3

HOG visualization 4

Figure 3:


Predictions Distribution
2757
92
9
7
2500
2000
1500
1000
500
0
Count
pure
broken
silkcut
discolored
Labels