Assignment 2

Objective:

Using the dataset collected decide on an experimental design, extract features of the group’s choice, train a probabilistic generative classifier model to classify the images.

Approach:

We decided to split our datasets so that 1/3 of our datasets would be used for testing and 2/3 would be used for training. This ratio was consistent with guidance we found online on how to divide a dataset for machine learning. We decided to use features related to the edges of the bricks. We assumed each of these features was normally distributed within in each class. From this we made a 5-varible Gaussian probabilistic function for each class using the five features extracted. For testing we extracted the features of the test data and input them into the probabilistic function we generated then returned the class that had the highest probability of having those features.

Questions:

1. What features did you select and implement?  Why did you select these features and how do you think they will be informative for this classification problem?

The features we selected were the percentage of the picture that was a vertical edge, the percentage of the picture that was a horizontal edge, the percentage of the picture that was an edge at 45°, and the percentage of the picture that was an edge at 135°. These features were selected because this is a common method to identify bricks with machine learning, the orientation of the edges of the bricks is how we as humans recognize which brick patterns are which, and we believe that other brick patterns and non-brick patterns images will be easy to distinguish from the rest using these features because they will have a completely different break down of types of edges than the brick patterns of class 1-3.

1. Given your data set and features, were there any outliers in your data set that needed to be removed or handled specially during pre-processing/feature extraction?  If so, what was causing these images to be outliers? If not, do you think your data set is a good representation for the problem or are you missing imagery of a certain type in your data set that could impact your overall performance?

Our dataset only had one example of class 0 images so although we were able to generate a probabilistic model for this class, we know that it cannot be accurate based on this extremely limited data. Also, we could not even test it against any other data.

1. How effective is your trained system (preprocessing + feature extraction + classifier)? Do you think you already have a system that would be competitive/effective for the project overall?  If so, why?  If not, what is needed to improve your approach?

It is hard to determine how successful our system is because the datasets which we trained and tested on were so small. This system will need to be trained over significantly more data before we can make a true assessment as to how well the system works. With that in mind, over our very limited dataset, performed well as shown in the confusion matrices below.

Final Observations:

The class which we had the most examples for, class 2, got the most correct in the testing, which is shown in our confusion matrix. This leads us to believe that given more datapoints for each class our model will improve.