In your report, please:
1. Describe the steps and parameters in your MED pipeline with two types of features: SURF and CNN. If you do the VLAD bonus, please also provide explanation about your VLAD implementation.
2. Report the confusion matrix for multi-class classification in your validation set. Which feature is better? Which class(es) is harder and with which it is confused? Do you have insight or explanation for the difference in accuracy?
3. Report the time your MED system takes (CPU time) for feature extraction and classification on the testing set. Please also tell us the amount of credits left on your AWS account.

My MED pipeline implementation is as follows:

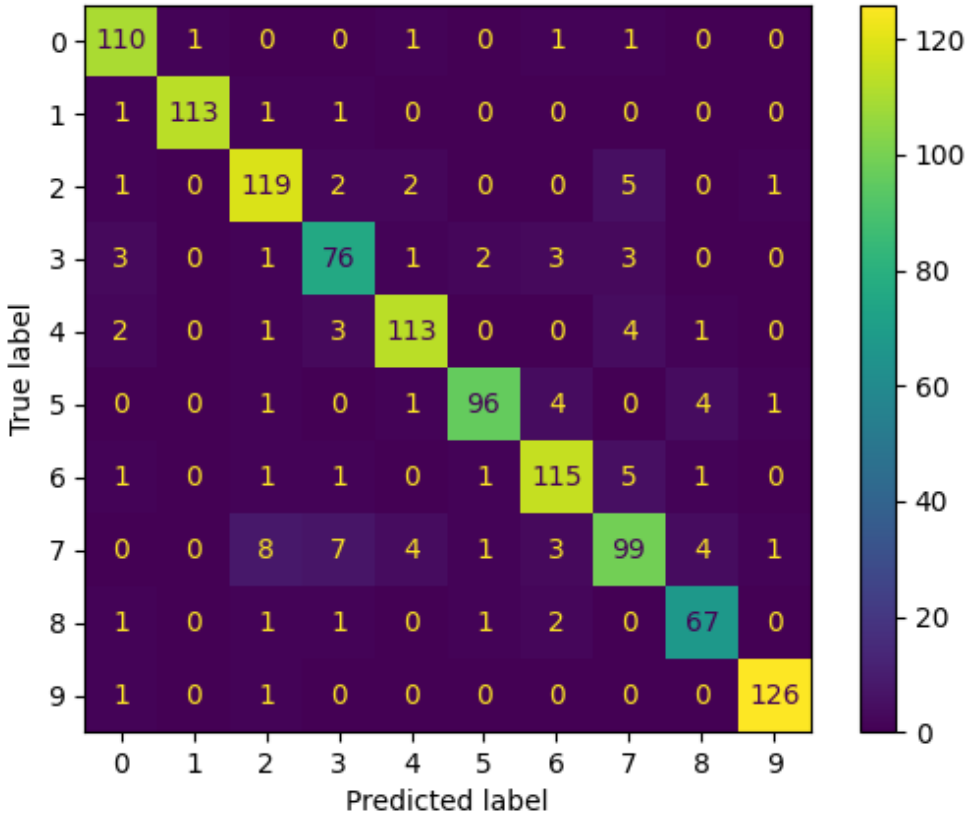
1) Extract the visual features from the videos using ffmpeg
2) In order to save process time for repetitive experimentation, I increased the size of the AWS volume storage attached to my instance from 100 GB to 300 GB to store all the images.
3) Extract the Speeded-Up Robust Features (SURFs) using OpenCV at a frame rate of 1.5.
4) To encode the Bag-of-Words representation, I trained the k-means clustering algorithm with the 1 percent of the entire SURF feature set. I tried with 256 cluster centers. Since the memory requirement exceeds the memory on board of a t2.large, I changed the instance type to a t3.2xlarge to accommodate the memory need.
5) To extract the CNN features, I completed the provided cnn_feat_extraction.py script to save the corresponding layers of some popular CNN architectures, such as ResNet18, ResNet50, and ResNet101. The entire CNN features extraction procedure takes around 5 hours to complete for each architecture.
6) Then I trained the classification model based on MLP.
7) I trained a few different models with different inputs, the details can be seen in the following chart.

Model	Input	Training Accuracy	Validation Accuracy
MLP (hidden_layer = (100,), relu activation, adam solver, 5000 max iteration)	SURF-256	99.04%	48.32%
MLP (hidden_layer = (100,), relu activation, adam solver, 5000 max iteration)	ResNet18-512	99.79%	87.34%
MLP (hidden_layer = (100,), relu activation, adam solver, 5000 max iteration)	ResNet50-2048	99.231%	89.724%
MLP (hidden_layer =	ResNet101-2048	100%	91.2621%

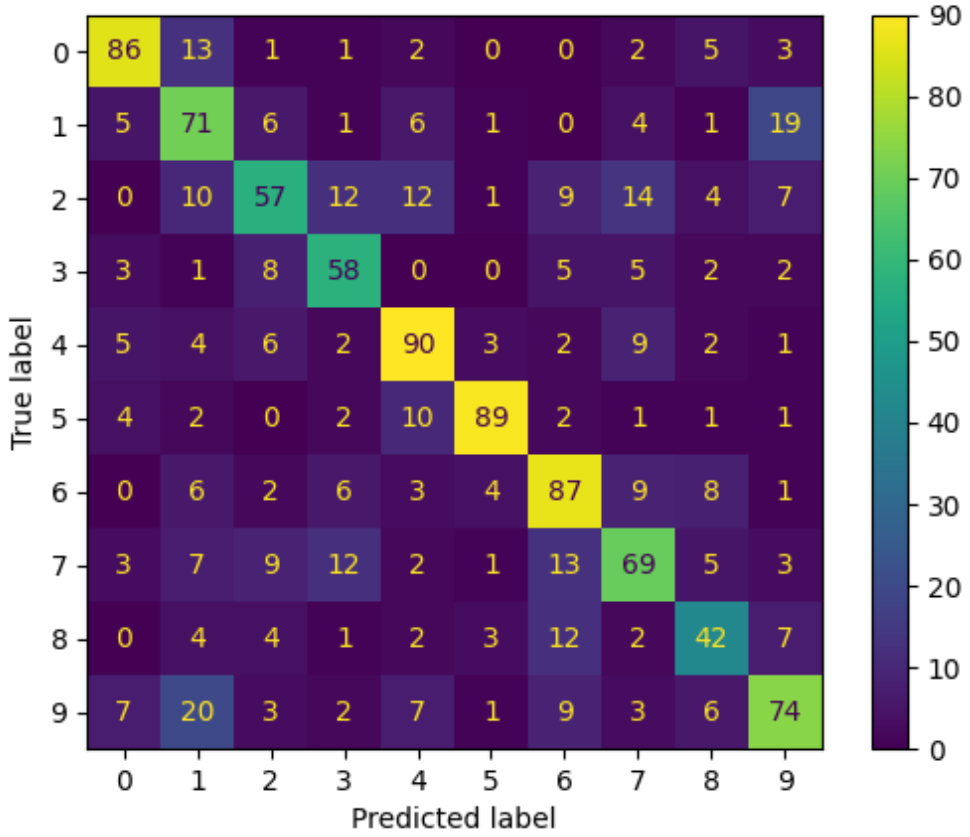
(100,), relu activation, adam solver, 5000 max iteration)			

8) I split the trainval set into an 80% training set and 20% validation set with shuffling. Then, I used 5-fold cross-validation to evaluate the mean top-1 accuracy and its standard deviation.

Results and Confusion Matrix:
The model with the highest top-1 accuracy that I trained is the MLP model with one hidden layer size of 100, ReLU activation, Adam solver, and 5000 max iterations. The confusion map is as follows:



The most difficult classes to classify using visual features are class 7 (singing) and class 2 (playing guitar), as well as class 3 (playing piano). This is probably because the movement of playing the instrument is hard to distinguish. Compared the confusion matrix with the one generated using sound features only, we can see a similar trend in misclassifying class 2 and class 7 as well.



The entire MED pipeline is composed of 4 parts: 1) extracting visual part 2) extracting SURFs 3) Use k-means clustering to extract Bag-of-Words SURF features 4) extracting CNN (ResNet in particular) features. 5) training the model 6) evaluation the model on the validation set.

1) extracting images (RGB features) takes around 6 hours. 2) extracting SURFs takes around 5 hours and 20 minutes. 3) extracting BoW features takes around 1 hour. 4) extracting CNN features takes around 5 hour for each architecture experimented using CPU only, around 15-18 hours in total. 5) reading the input data and training the model takes around 3-5 minutes 6) evaluating the performance takes around 15-30 seconds.

AWS Credits Left: I have used around 12 instance hours of t2.large ($0.09/hours) and 43 hours of t3.2xlarge ($0.3328), which sums up to be $15.3904 in total. I have $45.2-$15.39 = $29.81 left for the rest of the semester.