IMPLEMENTATION:

I. IDENTIFYING INSTRUMENTS IN VIDEO:

The task of identifying instruments in video or an image and localizing them fall in the domain of object detection and localisation. Object detection refers to the process of identifying and localizing certain ROI's within an image through the use of an image classification model trained on samples similar to that particular ROI.

Several algorithms have been developed that accomplish the same, and we have compared and contrasted among 3 of these algorithms, namely **YOLO, SSD, and Faster RCNN.**

We have decided to limit our comparison to these 3 algorithms in particular, due to their extensive documentation and proven success in solving problems similar to that of our's.

The below figure depicts when one must employ an algorithm belonging to the above list:

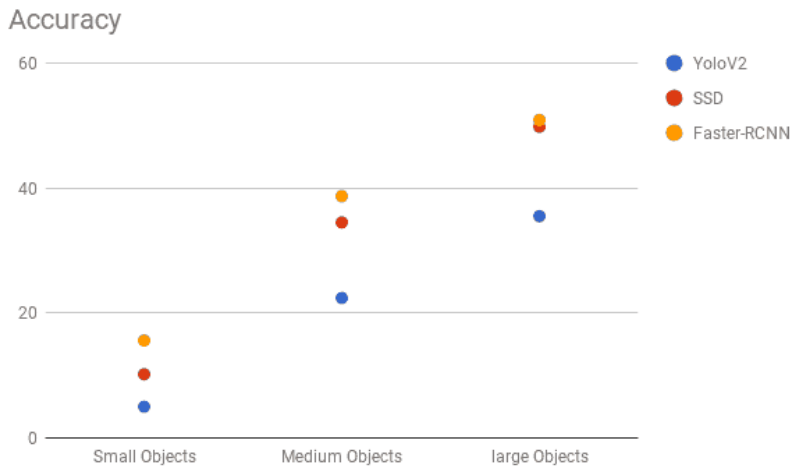

Fig.1. A comparison of YOLOv2,Faster RCNN and SSD in lieu of Accuracy and Object Size

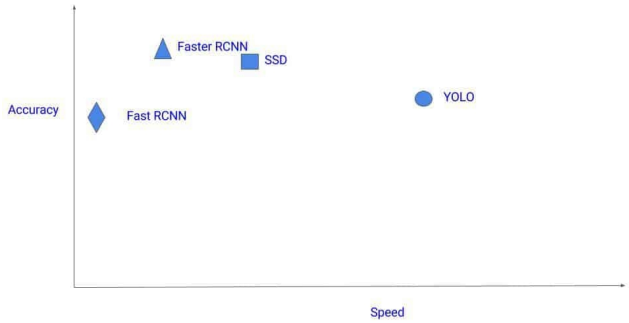

Fig.2. A comparison of YOLOv2,Faster RCNN, and SSD in lieu of Accuracy and Speed

II. FEATURE EXTRACTION

Once an instrument is detected and localized in a frame of the video, the probability of the detection is stored as a single value in a 7 dimensional feature vector. If the same instrument is detected once again in a future frame, the detection with the greater probability is stored instead of the one currently in the feature vector.

As a result, Each video can be converted into a 7 dimensional feature vector, which will be fed as input to our classification model.

Each value within the 7 dimensional feature vector corresponds to the probability of each of the 7 instruments that maybe in a video/frame.

III. DATASET CREATION

The dataset for the video classification model is in the .csv format, due to its compatibility with pandas and numpy. It consists of 51 rows and 8 columns, with each of the 7 columns representing the probability of a particular instrument being present in a video, and the 8^{th} column representing the label/genre of the video.

51 videos were collected and acted as input for our object detection program, that gave us a feature vector as output. As a result, 51 feature vectors were obtained and inserted into the .csv file.

IV. CLASSIFICATION MODEL AND RESULTS

The easiest way to select a classification algorithm is to trial as many as possible and evaluate the same, based on 3 different metrics, namely precison,recall and accuracy.

Precision attempts to answer what proportion of positive identifications were actually correct.
Recall attempts to anwser what percentage of total relevant results were correctly classified by our chosen algorithm.

We restrict our evaluation to Supervised classification algorithms, as our dataset is labeled in nature. Therefore, the following algorithms were considered:

1) SVM (Support Vector Machines)
2) MLP(Multi Layer Perceptron)
3) Random Forest

Our dataset would be well suited as input for a SVM,as it has proved to be particularly useful for datasets that are small, but have features that greatly discriminate between 2 classes.

We therefore, begun our evaluation with the SVM, and then tested the other 2 algorithms.
The results are shown as below:

Algorithm	Accuracy	Precision	Recall
SVM	0.8	1.0(hindustani) 0.67(carnatic)	0.67(hindustani) 1.0(carnatic)
MLP	0.8	1.0(hindustani) 0.67(carnatic)	0.67(hindustani) 1.0(carnatic)
Random Forest	0.8	1.0(hindustani) 0.67(carnatic)	0.67(hindustani) 1.0(carnatic)

Its evident from the above that all 3 algorithms gave near equal performance. However, the time taken to train the SVM was shorter than that of the MLP and Random Forest. Therefore, the SVM was chosen to be the best choice for the given classification problem.

V. CONCLUSION AND FUTURE WORK

In this paper, we have redefined the way in which we classify music videos, leveraging object detection as a means of feature extraction from music videos. We then make use of these vectors and train a SVM to classify the video into one of Hindustani or Carnatic genre.

We ,however, do not explore the possibilities of further increasing our classification accuracy,as we have ignored the audio perspective of the video. Longer feature vectors can be generated through the means of Spectrogram analysis and extracting significant audio features like pitch,timbre,tempo etc.

We also do not explore the possibility of leveraging human pose estimation as a means of music video classification.

The proposed work would be a stepping stone into solving Object based video classification problems in domains closely related to artistic performances.