International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Musical Instrument Recognition using CNN and SVM

Prabhjyot Singh[1], Dnyaneshwar Bachhav[2], Omkar Joshi[3], Nita Patil[4]

[1,2,3]Student, Computer Department, Datta Meghe College of Engineering, Maharashtra, India
[4]Professor, Computer Department, Datta Meghe College of Engineering, Maharashtra, India

---

Abstract - *Recognizing instruments in music tracks is a crucial problem. It helps in search indexing as it will be faster to tag music and also for knowing instruments used in music tracks can be very helpful towards genre detection (e.g., an electric guitar heavily indicates that a track is not classical). This is a classification problem which will be taken care of by using a CNN (Convolutional Neural Network) and SVM (Support Vector Machines). The audio excerpts used for training will be preprocessed into images (visual representation of frequencies in sound). CNN's have been used to achieve high accuracy in image processing tasks. The goal is to achieve this high accuracy and apply it to music successfully. This will make the organization of abundant digital music that is produced nowadays easier and efficient and helpful in the growing field of Music Information Retrieval (MIR). Using both CNN and SVM to recognize the instrument and taking their weighted averages will result in higher accuracy.*

Key Words: *CNN, SVM, MFCC. Kernel, Dataset, Features*

1. INTRODUCTION

Music is made by using various instruments and vocals from a human in most cases. These instruments are hard to classify as many of them have a very minor difference in their sound signature. It is hard even for humans some time to differentiate between similar sounding instruments. Instrument sounds played by different artists are also different as they have their own style and character and also depends on the quality of the instrument. Recognizing which instrument is playing in a music will help to make recommending songs to a user more accurate as by knowing persons listening habit we can check if they prefer a particular instrument and make better suggestions to them. These suggestions could be tailored to a persons liking. Searching for music will also be benefited as for the longest time only way songs or music has been searched is by typing name of a song or an instrument, this can be changed as we can add more filters to the search and fine tune the parameters based on instruments. A person searching for an artist can easily filter songs based on instruments playing in the background. It can also help to check which instruments are more popular and where this will help to cater more precisely to a specific portion of the populous. Using CNNs to classify images has been extensively researched and has produced highly accurate results. Using these detailed results and applying them to classify instruments is our goal. Ample research has also been carried out to classify instruments without the use of CNNs. Combining the results from CNN classification with classification by SVM to get better results is the eventual goal. The weighted averages of the predictions will be used. SVM have been used for classifications for a long time with great results and to apply these to instrument recognition in music is the key idea which may help to achieve satisfactory results. Every audio file is converted to an image format by feature extraction using LibROSA python package, melspectrogram and mfcc are to be used.

2. LITERATURE SURVEY

Previous research on this subject has been carried out with varied results and using methods ranging from Hidden Markov Model (HMM) to ANN and various others. Lewis Guignard et al [1] utilize machine learning techniques to construct a classifier for automatic instrument recognition given a mono-track audio recording. It focuses on on the classification of eight instruments commonly used in rock music bands. Audio samples are obtained from online and live recordings. Features are extracted using DFT. Several models were applied to the dataset, first in an exploratory manner K-Means, Principal Component Analysis (PCA) for instrument classification, a Support Vector Machine implementing a variety of kernels is used. For regularization, it used 10-fold cross validation (CV) to compare different models used. 10-fold cross validation trains on 90 % of the data and tests on the remaining 10 %, and repeats for all 10 ways to divide the dataset in this way. SVM performed better than PCA and K-Means.

Toni Heittola et al. [2] also proposes a novel approach to musical instrument recognition in polyphonic audio signals by using a source-filter model and an augmented non-negative matrix factorization algorithm for sound separation. In the recognition, Mel-frequency cepstral coefficients are used to represent the coarse shape of the power spectrum of sound sources and Gaussian mixture models are used to model instrument-conditional densities of the extracted features. The method is evaluated with polyphonic signals, randomly generated from 19 instrument classes. GMM is used as a classifier

here. Nineteen instrument classes are selected for the evaluations of the following instruments are accordion, bassoon, clarinet, contrabass, electric bass, electric guitar, electric piano, flute, guitar, harmonica, horn, oboe, piano piccolo, recorder, saxophone, trombone, trumpet, tuba. The instrument instances are randomized into training (70%) or testing (30%) set. This method gives good results when classifying into 19 instrument classes and with the high polyphony signals, implying a robust separation even with more complex signals.

Akram Azarloo et al [3] proposed a taxonomy of musical ensembles, which is automatically built, to represent every possible combination of instruments likely to be played simultaneously. Trumpet, Percussion, Drums, Piano, Double bass, Guitar are used for experimentation also female and male singing voices are considered as possible "instruments". MLP and K-Nearest Neighbors (K-NN) are the classifiers used in this project. Features proposed in the project are of 4 different categories: temporal, cepstral, spectral, and perceptual. Features. UTA algorithm is used for calculating average of one feature in all instances. The best of accuracy average is in the case of MLP classifier as compared to K-NN classifier. The usefulness of a wide selection of features and after feature selection application with UTA algorithm accuracy average increased in the K-NN classifier.

3. METHODOLOGY

The model proposed in this project is a combination of Convolutional Neural Network and Support Vector Machine. Figure 1 shows the proposed System Architecture.

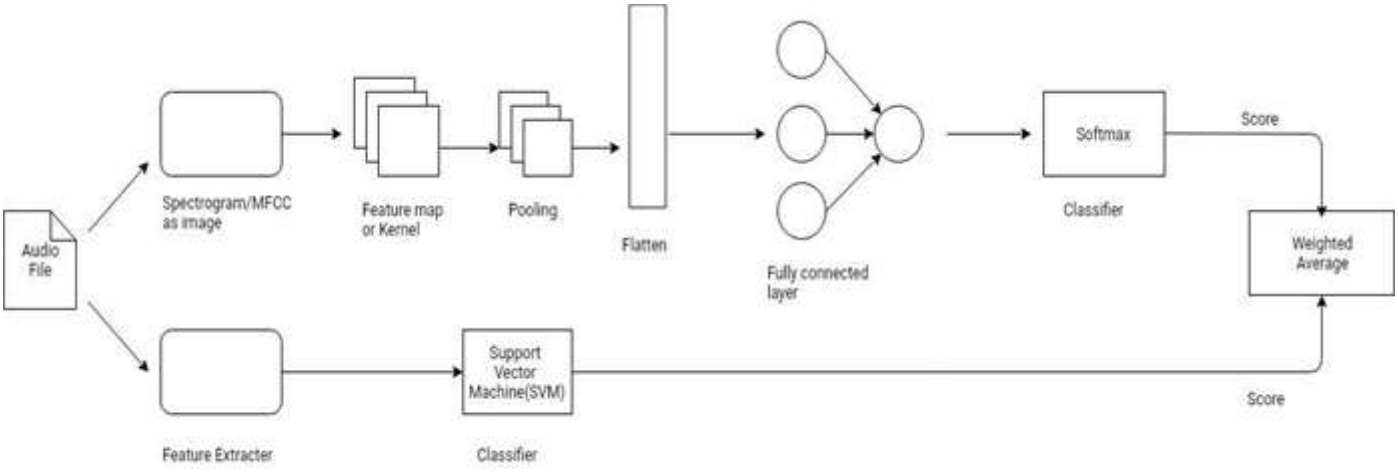

Fig-1: System Architecture

The model consists of four convolutional layers, followed by one fully connected layer and one softmax layer and SVM. It has been shown that the deeper the convolutional layer is, the more abstract the features it learns. MFCC is used for feature extraction for SVM. The results obtained from both the CNN and SVM are added to get the weighted average, which will give the better performance in terms of instrument identification.

3.1 CNN

Convolutional Neural Networks are similar to Neural Networks as they have weights which change overtime to improve learning. The big difference is that CNNs are 3-D as they are images and hence are different from ANNs which work on 2-D vectors. CNNs are pretty straightforward to understand if previous knowledge of ANN is present. CNNs use the kernel also known as feature map or a filter. This filter is slided over the input image and as it moves block by block we take the cross product of the filter with the image. Filter moves over the entire image and this process is repeated. As we can see in the there is not just a single filter but lots of them depending on the application or model. Here we show a single set of convolution layer i.e. bunch of filters but there are usually multiple convolution layers in between. And as we progress through multiple layers the output of the previous layers become the input for the next layer, and we again perform cross product to get the finer details and helps the model learn more in-depth. It takes a lot of computing power to handle this much data as images in real world have lots of details and have higher size hence increasing the size of each filter and computation required.

To overcome this we pool these filters. Pooling is basically reducing the parameters of each filter. This is also called as down-sampling or sub-sampling. It works on each filter individually. Reducing the size improves the performance of the network. The most used technique for pooling is Max Pooling. In Max Pooling we basically again move a window over the filter but instead of taking a cross product, just select the maximum value present. It also has other advantages such as making influence of orientation change and scaling of image lesser. Orientation and scaling can make a lot of difference if not taken into consideration.

Flattening is the intermediate between convolution layers and fully connected layers. As we have 3-D data after pooling we convert it to 1-D vector for the fully connected layer. This process is done by flatten.

Fully connected layer is the one in which the actual classification happens. We extract features from the image in the Convolution part. This part is like a simple ANN which takes the features from the image as an input and adjust the weights based on back propagation.

CNNs are have been researched a lot in the past few years with amazing results. $ Karen Simonyan and Andrew Zisserman of the University of Oxford created a 19 layer CNN model more commonly known as VGG Net. It uses 3x3 filters, pad of 1 and 2X2 max pooling layers, with stride 2.

Google Net was one of the first CNN architecture which didn't just stack convolution and pooling layers on top of each other. It had an error rate of only 6.67%. They did not just add more layers to the model as it increases the computation power.

3.2 SVM

Support Vector Machines (SVMs) are one of the most popular strong machine learning algorithms for classification, for their ease of use and wide success. SVM is one of the efficient algorithm in machine learning. It is efficient in computation and robust in high dimension. It can be used for classification as well as regression problems. The architecture is very much simple. Number of features required is less for recognition. The aim of the SVM is to find hyperplane which involves multiple features that distinctly classifies the classes. Future data points or classes can be classified more accurately and confidently based on the optimal result of margin distance.[4]

Support vector machine is trained with dataset of musical instrument. While training the SVM feature extractor extract the features from the given input in the form of the feature set. Then the feature set get the basic information of every input. These feature sets classify the entire feature unit.[5]

4. FEATURE SELECTION

Identifying the components of the audio signal is important. So feature extraction is used to identify the basic important data of given input and it also removes the unnecessary data which comes in the format of emotion, noise, etc.

MFCC i.e Mel Frequency Cepstral Coeficients is a feature used in automatic speech and speaker recognition. Standard 32-ms windows is going to use for the extraction of features.

In this section our goal is to achieve a smaller feature set which will give us the quick result i.e recognizing the input in real-time. It will also not compromise the recognition rate.

5. DATASET

IRMAS dataset is divided into Training and Testing data. Audio files are sampled at 44.1 kHz in 16 but stereo wav format. Total of 9,579 audio files are present. The instruments present for recognition are cello (cel), clarinet (cla), flute (flu), acoustic guitar (gac), electric guitar (gel), organ (org), piano (pia), saxophone (sax), trumpet (tru), violin (vio), and human singing voice (voi). The number of files per category are shown in the brackets cel(388), cla(505), flu(451), gac(637), gel(760), org(682), pia(721), sax(626), tru(577), vio(580), voi(778). Additionally, some of the files have annotation in the filename regarding the presence ([dru]) or non presence([nod]) of drums, and the musical genre: country-folk ([cou_fol]), classical ([cla]), pop-rock ([pop-roc]), latin-soul ([lat-sou]).

The dataset is very well constructed by taking into consideration change of styles in the past decades, also different ways in which people play the instrument and the production values associated. In the training dataset 3 seconds excerpts are present out of 2000 distinct recordings.

The training set contains 6705 audio files and 2874 audio files are kept for testing. The length of the sound excerpt for testing is between 5-20 seconds.

6. CONCLUSION

To achieve better accuracy then previous researches.

REFERENCES

[1] M. Young, the Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Toni Heittola, Anssi Klapuri and Tuomas Virtanen,"Musical Instrument Recognition in Polyphonic Audio Using source filter model for sound separation", Department of Signal Processing, Tampere University of Technology,2009 International Society for Music Information Retrieval.

[3] Akram Azarloo, Fardad Farokhi,"Automatic Musical Instrument Recognition Using K-NN and MLP Neural Networks", Faculty of Electrical and Electronics Engineering Islamic Azad University,2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks.

[4] Philippe Hamel, Sean Wood and Douglas Eck,"Automatic Identification of instrument classes in Polyphonic and Poly-Instrument Audio",Departement d'informatique et de recherche opérationnelle, Université de Montréal,2009 International Society for Music Information Retrieval.

[5] Swati D. Patil1 and Tareek M. Pattewar,"Musical Instrument Identification Using SVM and MLP with Formal Concept Analysis",1PG Student, Department of Computer Engineering, SES's R. C. Patel Institute of Technology, Shirpur, Maharashtra, India, 2 Assistant Professor, Department of Information Technology, SES's R. C. Patel Institute of Technology,2015 IEEE.