MULTI-MODAL PIANO NOTE DETECTION USING AUDIO AND VIDEO

by

NIRMALKUMAR LAXMANBHAI PATEL

A project report submitted in conformity with the requirements
for the degree of Master of Science, Information Technology

Department of Mathematical and Physical Sciences
Faculty of Graduate Studies
Concordia University of Edmonton

MULTI-MODAL PIANO NOTE DETECTION USING AUDIO AND VIDEO

NIRMALKUMAR LAXMANBHAI PATEL

Approved:

Supervisor : Dr. Baidya Nath Saha	Date
Committee Member	Date
Dean of Graduate Studies: Dr. Alison Yacyshyn	Date

MULTI-MODAL PIANO NOTE DETECTION USING AUDIO AND VIDEO

NIRMALKUMAR LAXMANBHAI PATEL
Master of Science, Information Technology

Department of Mathematical and Physical Sciences
Concordia University of Edmonton
2022

Abstract

Many people have been interested in music recognition. The automated transcription of musical compositions and the identification of sound sources, such as the sort of instruments used, have taken a lot of time and work. With the rise of personal computers and multimedia systems in recent years, research in these areas has gotten a lot of attention. In our paper, we have chosen a piano based song for the purpose of analysis. We have divided the song in chunks called frames for note recognition. Initially, we performed manual analysis to recognize the notes so that we have the correct notes. Then after, we have used finder tip following technique for tracking the notes which are played. This is our input dataset for image or frame based input. Subsequently, the audio is extracted and divided to chunks similar to number of frames in the video. We have performed audio frequency analysis to perform note detection based on the audio. When the variables of interest can't be measured directly but an indirect measurement is available, Kalman filter and particle filter are used to estimate them as best as possible. They're also used to obtain the best approximation of states in the presence of noise by integrating readings from numerous sensors. The novelty of our research is that we have implemented Kalman filter and particle filter based on audio and video based input instead of sensor data which is never used before.

Keywords: Kalman filter, Machine learning, Image processing, Audio processing, Note detection, Particle filter, Pitch detection

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Personal digital technology has revolutionised how music is distributed and stored, and thus increased interest in and focus on how information technology may be used to create this type of material. Nowadays, computers play a significant role in nearly every element of music creation, from exploring personal collections to discovering new musicians to controlling and defending the rights of music composers. The US remains the most lucrative market for digital music, with recent years seeing a billion-dollar industry for digital music sales. People now have access to millions of music clips from online music applications through their smartphones thanks to the digital revolution. Therefore, most individuals wish to learn "How to play it" on their own using music books, scripts, or music notation. For a skilled listener, extracting the necessary musical information from a live or recorded performance is very simple, but for a student and computer, it is quite challenging. It would be preferable to collect this information, such as piano notes, quickly, accurately, and automatically for a variety of practical applications. The major goal of this project is to develop a fusion algorithm that takes the audio and visual signal as input and produces a learning note sheet for musicians, producers, composers, DJs, remixers, teachers, and music students.

The process of automatically converting a musical audio input into a symbolic representation, such as a piano roll or music score, is known as automatic music transcription (AMT). It has several uses in music composition, music information retrieval, music education, and music visualisation.

The Kalman filter is a powerful mathematical tool that is becoming more and more crucial in a variety of applications, including parametric estimating and computer graphics. The good news is that understanding and using Kalman filters doesn't require being an expert in mathematics. Despite the Kalman filter's approximately

30-year history, it has just lately begun to appear in a wide range of applications. For a sizable class of problems, the Kalman Filter is the best estimator that can be used, and it is also a very good and valuable estimator for an even larger class. The age-old problem of how to extract accurate information from faulty data is addressed via Kalman filtering. More urgently, how can you revise your "best" assessment of the condition of a system if fresh, unreliable data keep coming in? The Kalman filter is intended to remove undesired noise from a stream of data, much like a coffee filter keeps undesirable grounds out of your morning mug. There are countless uses for Kalman filters. A few applications of Kalman filtering include satellite orbit determination, radar tracking, sonar range, and navigational and guiding systems. Thousands of other articles on filtering-related topics and applications have been written as a result of Kalman and Bucy's initial publications. Their work has encouraged more mathematical study in fields like numerical linear algebra approaches.[1]

Early in the 1990s, the particle filter gained popularity, and estimate issues have been resolved ever since. The extensive availability of lesson content and code examples makes it simple to comprehend and use the standard algorithm. The conventional particle filter method has undergone much study to enhance performance and application in numerous ways in the next years. Therefore, choosing and putting into practise an advanced version of the particle filter that goes above and beyond the basic algorithm and suits a particular estimate issue needs either a profound comprehension or a thorough literature research. The latter can take a lot of time, especially for people with little practical experience. The lack of implementation information in theory-focused studies makes this effort more difficult.

1.2 Problem Statement

Piano music detection is an interesting topic for past few years. Plethora of researchers have worked on detecting musical chords using various methods which uses Pitch, frequency and noise removal techniques. However, unwanted sound "noise" will defer the notes by adding pitch or frequency in the music. So, proposed paper consists the techniques to verify the audio detected notes using video or frame detected notes so that we can improve the efficiency of right notes.

Music detection and recognition becomes hard when there is a noise in it. So, noise can differ the resultant note or recognized note by adding or differing the pitch and/or frequency. Here, we are going to use multi modal based approach to derive audio and video based note which uses various signal processing algorithms such as Kalman filter, particle filter based audio & video based note detection and recognition.

1.3 Contribution of the Thesis

- The research work involves the collection, cleaning, and analysis of data for the extraction of useful insights/information. The contribution to this project work involved the collection of video of selected music and piano demonstration of the song.
- The video is converted into equivalent frames which is used as data for detection and recognition.
- The Audio is extracted and converted to wav form for simplification. The audio signal is sliced with identical number of the frames in the video to derive the note in each frame.
- With manual efforts the notes played for the selected music are collected in text form to use later for verification purpose.
- With Hand recognition technique and finger following method, we have analysed and detected the notes played for the duration of audio.

1.4 Organization of the Thesis

For this research work, there are seven (7) chapters. Chapter 1 is the Introduction which provides a detailed background of the research work, explaining the reasons for research work, the problem statement, the detection and recognition of Piano notes for selected music, and how the different models are used for it. Chapter 2 carries the Literature Review and is sectioned into nine various important parts for musical note analysis and recognition. It starts with discussing the audio signal classification followed by musical note segmentation and audio signal analysis. Then after, Several features of audio are discussed such as Tonality and segmentation, pitch detection and polyphonic Transcription. In the last three parts, the method of finger tip detection and two multi modal based approaches for musical note recognition. Chapter 3, which is the methodology discusses in detail the audio processing methods that are used for the research purpose. Chapter 4 is the video based note processing discussion of various methods that are attempted during the findings. Chapter 5 consists of Kalman filter and particle filter based multi modal approach of solving the problem of note detection and recognition. Chapter 6 talked about the results and discussions of the analysis. Chapter 7 concludes this research work and also discussed the future works for further analysis.

Chapter 2

Literature Review

Music detection and Image detection are the areas of attraction for researchers from a log time. Here, we will see what researches has been done and the techniques are used to improve it.

2.1 Audio Signal Classification

A system for classifying audio signals analyses the input audio signal and generates a label for the output signal. These are used to categorise speech and musical signals. The classification may be done based on the pitch, musical style, musical pace, and musical rhythm. The signal classifier examines the audio format's content in order to extract details about the content from the audio data. This is also known as audio content analysis, which also includes signal-based content information retrieval. This study presents the audio signal classification's implementation. Pitch, timbral, and rhythmic characteristics have all been explored in relation to their capacity to discriminate between various audio formats. Both the key feature selection process and the typical classification methods have been described. Finally, a method known as the confusion matrix has been researched in order to assess how well the categorization system performs.[2]

In order to determine which of a number of classes a sound is most likely to belong to, audio signal classification (ASC) involves extracting pertinent information from a sound and utilising those features to categorise the sound. Depending on the application's classification domain, several feature extraction and grouping techniques may be employed. This article provides basic information on signal processing, spectral analysis, psychoacoustics, and auditory scene analysis that is important to comprehend the overall study topic of ASC[3].

The fundamental components of categorization systems are also covered. We also address clustering techniques, analysis time, and perceptual and physical aspects. ASC is studied in relation to concealed Markov models and neural networks. An summary of the ASC's current situation is provided with the presentation of these strategies[3].

2.2 Musical Note Segmentation

In this study, *Gordana et al.*[4] provide a method for musical stream note segmentation that combines time-domain and frequency-domain analytic techniques. The technique has two clear advantages: first, it produces findings that can be trusted with a very low likelihood of either missing or incorrectly identifying notes, and second, it has a temporal precision of about 1 msec. They have used the technique on recordings of a range of monophonic musical instruments, including as the clarinet, piano, and violin, with results that range from 95% to 100% accuracy[4].

2.3 Audio Signal Analysis

The solution by *Geoffroy et al.*[5] deals with the automated estimate of a music track's key (key-note and mode) from the analysis of its audio data. Such a system often relies on a series of steps, each of which formulates an assumption about the signal content or the musical content: spectral representation, mapping to chroma, and determination of the overall key of the musical composition. Here, they explore the underlying theories, evaluate them, and suggest advancements above the state of the art. They specifically suggest using a Harmonic Peak Subtraction algorithm as the system's front-end and assess the effectiveness of a hidden Markov model-based method. they then evaluate our strategy against other strategies utilising a database of 302 baroque, romantic and classic music tracks[5].

An article by *Chunghsin Yeh, et al.*[6] provides an STFT (short-time Fourier transform) representation-based frame-based method for estimating multiple fundamental frequencies (F0s) of polyphonic music sources. It is suggested that in order to estimate the number of sources and their F0s, the noise level be estimated first. Following that, all potential combinations of the chosen F0 candidates are then jointly evaluated. Their hypothetical partial sequences are constructed from a set of F0 hypotheses, taking into account any potential partial overlap. The sets of F0 hypotheses that are most plausible are chosen using a scoring function. Hypothetical sources are gradually blended and repeatedly verified in order to deduce the ideal combination. When the overlapping partials are taken into account, a hypothetical source is deemed to be legitimate if it either explains more energy than the noise or considerably improves the envelope smoothness[6]. The suggested system was entered into the 2007 and 2008 MIREX (Music Information Retrieval Evaluation eXchange) competitions, where the accuracy was assessed in relation to the number of sources inferred and the accuracy of the F0s computed. The favourable outcomes illustrate its competitive performance when compared to other cutting-edge techniques[6].

2.4 Tonality and Segmentation

A key component of musical structure is tonality. It explains the connections between the harmony and melody components. One of the primary challenges in tonal analysis is key detection in music. Automating the examination of the evolution of musical themes and emotion will need the creation of computational models that imitate perception and key detection. Practically speaking, intelligent music editing systems and automatic indexing of music repositories will benefit from semantic segmentation of music, including segmentation based on key change. Additionally, recognising recurring patterns in music is essential for music indexing and searching, and key identification is a crucial stage in this process [7].

2.5 Pitch Detection

Ghiurutan et al.[8] classify noises into musical notes in their study using supervised learning techniques as Support Vector Machine, Stochastic Gradient Descent Classifier, and K-Nearest Neighbors algorithms. The musical note identification problem, or, alternatively, the pitch detection problem, can be stated as follows in the context of supervised machine learning: given a musical piece that has been broken up into a certain number of time frames, categorise each of those time frames into one of various classes of pitches. Following the categorization procedure, the musical note for each segment of the composition will be known since each pitch correlates to a single musical note individually, and as a result, the musical score will be fully established[8].

2.6 Polyphonic Transcription

The process of turning an acoustical waveform into a parametric representation, where notes, their pitches, beginning times, and duration are retrieved from the waveform, is known as polyphonic pitch identification. It is not innate in how people perceive music, and transcription is a challenging cognitive activity. Additionally, it is a particularly challenging issue for modern computer systems. Robust algorithms with performance that should deteriorate smoothly as noise levels rise are needed to separate notes from a combination of other sounds, which may include notes produced by the same or other instruments or just background noise [9].

2.7 Mediapipe Method

Using a stochastic linear formal grammar module is the Mediapipe technique. Through a process of feed-forward grammar checking, this module successfully improves the classification accuracy of the machine learning model that classifies input hand gesture sequences from photos. Based on the many gesture combinations that may be produced, the system has a default, preset syntax. There is still hope that some sort of post processing of predicted gestures within a sequence generated by a machine learning system may improve accuracy, despite the fact that this may be more challenging to scale to hand gesture sequences like finger spelling, where there are many more possible sequences.[10]

2.8 Multi-object Tracking- Kalman Filter

Keeping track of several targets while maintaining their identities is crucial in some applications, including behaviour comprehension. However, owing to various real-time circumstances, tracking results could be unsatisfactory. These situations, which may be seen while objects are being monitored in real-time, include inter-object occlusion, occlusion of the objects by backdrop obstructions, splits, and merges. This work proposes a feature-based Kalman filter motion method for handling multiple object tracking. The initialization of tracking is entirely automated and doesn't involve any form of user input. The difficulties of correspondence after a split may be resolved by creating a Kalman filter motion model with the characteristics centroid and area of moving objects in a single fixed camera monitoring scene, utilising information collected through detection to determine whether a merging or split occurred. The suggested method has been tested on a series of images of people and vehicles. The outcomes demonstrate that the suggested method successfully tracks many moving objects in perplexing conditions[11].

2.9 Tracking- Particle Filter

A novel technique for performing target recognition and tracking in noisy environments is presented by Y. Boers et al. Instead of using conventional metrics to do tracking, they instead use the unprocessed radar video data. Following that, detection is dependent on a posteriori information, or the probability density of the state given these prior observations (in this case video data). For obvious reasons, this method of data processing and tracking is also known as track before detect (TBD). The fact that this methodology uses integrated and kinematically linked energy to determine whether or not a target is present gives it an edge over traditional tracking. The proposed technique outperforms the conventional method for tracking weak objects in noise.A brand-new technique for locating and following a potential target amid noise is presented by Y. Boers et al. They track objects using the raw radar video data rather than conventional measures. Then, detection is predicated on the a posteriori information, i.e., the state's probability density given these prior observations (in this case video data). Track before detect (TBD) is another name for this kind of data processing and tracking, which makes sense. This method's benefit over traditional tracking is that it bases its determination of whether a target is present or not on integrated and kinematically linked energy. The proposed approach performs weak target tracking in noise better than the conventional method[12].

Chapter 3

Note Detection using Audio

Musical signals are a broad and complicated collection that has only just begun to receive the attention it deserves among the various signals attracting the attention of the signal processing community. For thousands of years, music has been a part of human civilization, and numerous authors have written about its strange power of expression. To comprehend what is included by the single term music is equivalent to comprehending the vastness of a galaxy. This expression has taken shape in so many genres, with so many instruments, voices, and combinations thereof, communicating so many emotions. Many signals, such as sonar, radar, and communications signals, are built with autonomous signal processing in mind. Musical signals, like speaking signals, have been constructed primarily with the human ear in mind throughout their history. This immediately raises the stakes, as the brain possesses exceptional pattern recognition and context-awareness abilities. To distinguish speech from music, despite the huge number of literary art, speech appears to be primarily an aesthetic medium, whereas music appears to be primarily an information medium. Musical lines employed in the military, for example, where a bugle may announce a muster, a charge, or a retreat, and a drum might help a unit's planned march, are notable outliers.

3.1 Audio Extraction

In the proposed solution, we are going to use audio and video as inputs to the model. So, it is quite important to extract the audio from the chosen video. Audio extraction is arduous task to perform, we have used onset and offset based audio segmentation algorithm proposed by *Hu and wang*[13]. The wrapper function provided by Pydub[14] library takes time as input and slice the audio accordingly with out loosing the quality of the song. The video has total 851 frames so we had sliced audio in 851 parts. As shown in figure 3.1, various audio signal presentation techniques are used to find out the meaningful information such as frequency distribution to calculate the maximum repeating frequency.

The constant-Q transform[15], abbreviated CQT in mathematics and signal processing, converts a data series to the frequency domain. It has connections to both the Fourier transform and the intricate Morlet wavelet transform[16]. Its structure lends itself to musical depiction. The transform may be conceptualised as a sequence of filters, with each filter having a frequency that is logarithmic spaced apart and a spectral width that is a multiple of the filter before it:

$$\delta f_k = 2^{1/n} \cdot \delta f_{k-1} = \left(2^{1/n}\right)^k \cdot \delta f_{\min},$$

In the proposed solution, the audio segmentation technique is used to slice the audio waveform to the identical number of frames in the video, this technique lead us to derive notes from audio as well as video at a frame. To process the audio, we had used a technique to find out the frequency spectrum and the frequency magnitude of the each audio frame. For particular frame, there is a function which predict the range of frequency in which the note is captured. The next step is to find out the maximum repeating frequency and according notes from the pre-define note - frequency list.

As can be seen in the figure 3.5, the n^{th} frame is correlated with the n^{th} audio slice. Both will be used in the process of deriving note.

3.2 Universal Frequencies

The length and tension of a piano string determine the frequency at which it vibrates, and by varying the tension, we can control the precise frequency at which each string vibrates. The same idea holds true for all other musical instruments, including drums. So, there is a frequency that corresponds to each note on the piano keyboard. For instance, the note G2 on a piano is at 98 Hz, the note A2 is at 110 Hz, and the note C3 is at 130.8 Hz. Every note on the piano between 50 and 400 Hz is depicted in the figure 3.7, along with the corresponding frequency value for each note[17].

Now, Each of the slices are evaluated by the frequency analysis and related functions which can determine the repeated frequencies and corresponding notes from the universal note frequency's dictionary. The resultant signal and information can be seen in the figure 3.1.

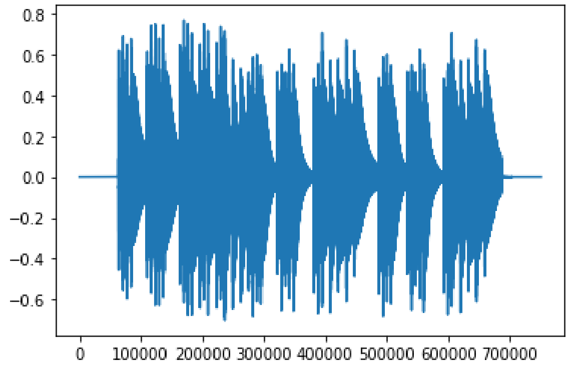


Figure 3.1: Original Audio Frequency Distribution

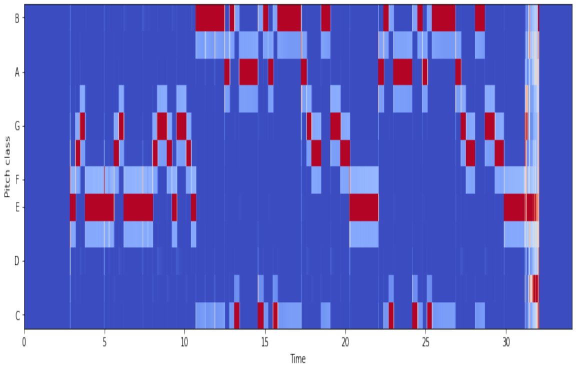


Figure 3.2: Constant-Q Power Spectrum

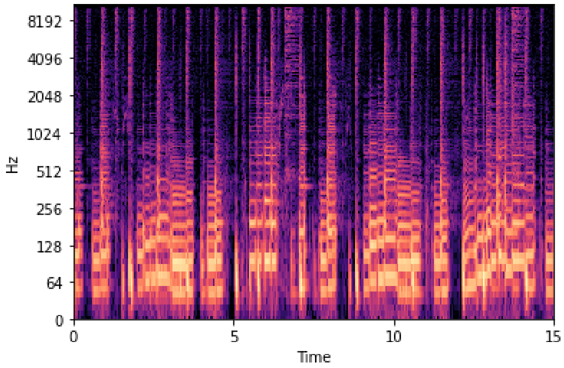


Figure 3.3: Linear Frequency Power Spectrum

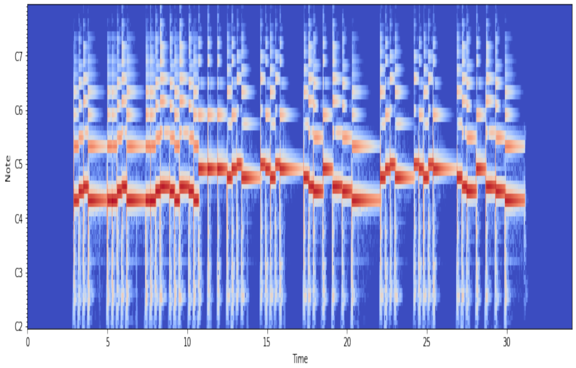


Figure 3.4: Short Time Fourier Transform

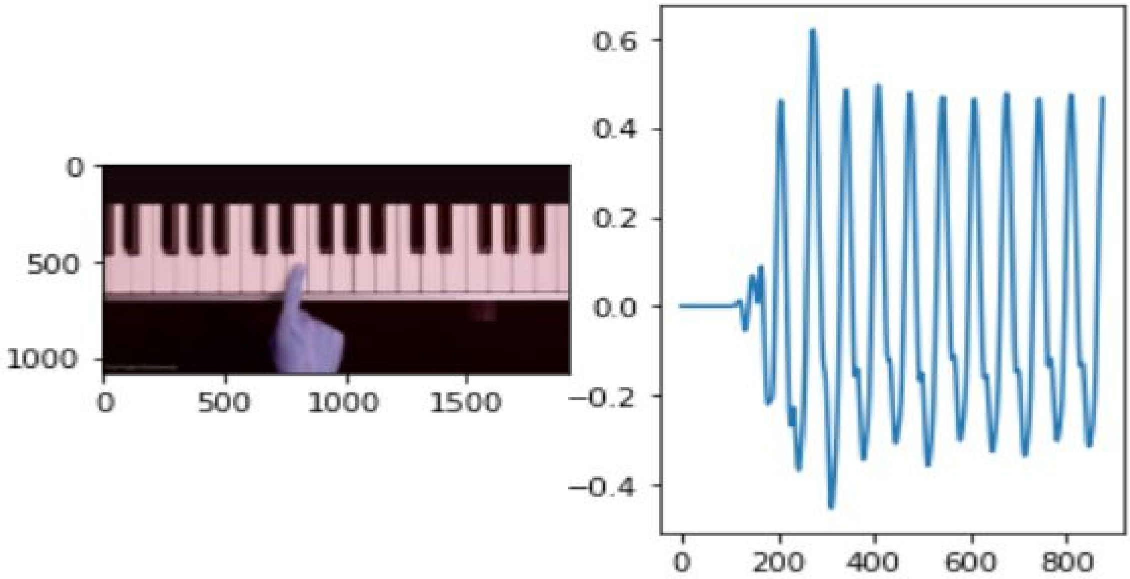


Figure 3.5: Audio slice and corresponding video frame

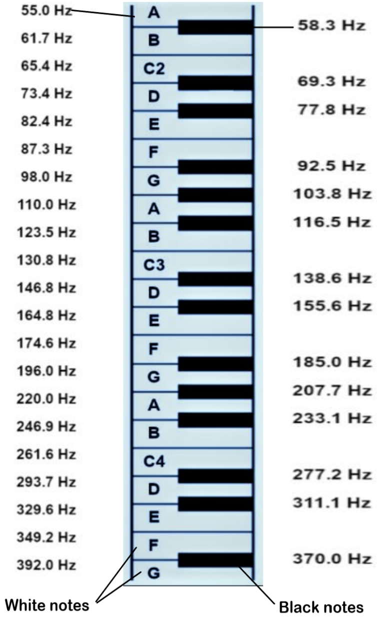
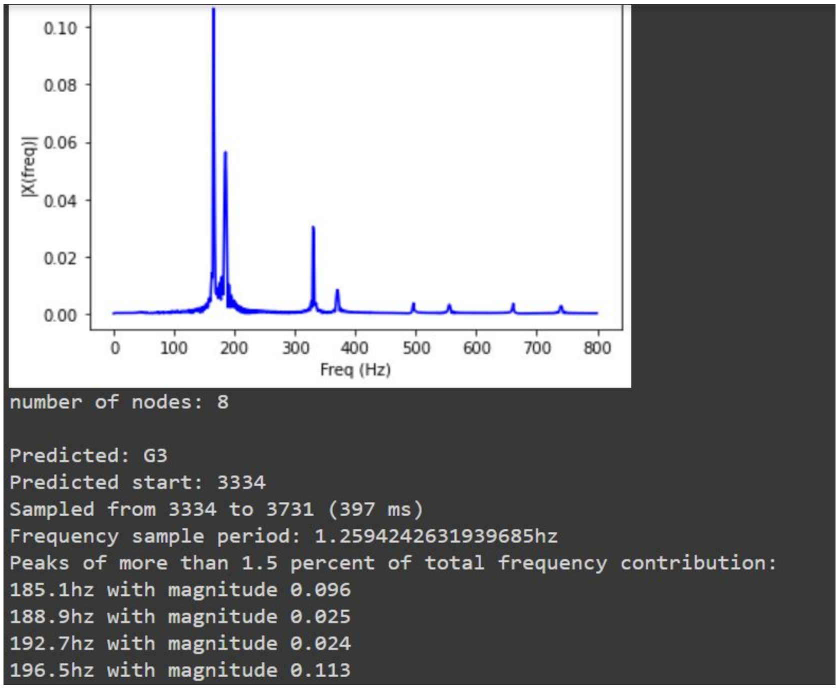


Figure 3.6: Note and frequencies



```
number of nodes: 8

Predicted: G3
Predicted start: 3334
Sampled from 3334 to 3731 (397 ms)
Frequency sample period: 1.2594242631939685hz
Peaks of more than 1.5 percent of total frequency contribution:
185.1hz with magnitude 0.096
188.9hz with magnitude 0.025
192.7hz with magnitude 0.024
196.5hz with magnitude 0.113
```

Figure 3.7: Resultant note and prediction information

Chapter 4

Note Detection using Video

4.1 Tile Bound of Piano Note

The localization of each note is important in proposed solution in order to match the finger tip location to find the corresponding note. The first step to detect the node successfully is to bound the area of each notes that are showing in through out the video. To achieve this goal, we had started with several techniques to define boundaries by various methods and detection. Among which we found that Manually defining the boundary of each note is way more fruitful than using any algorithms as the result of algorithms affects when the finger comes in the frame. So, each of the nodes has been noted manually and stored all the location of the white and black notes in the file. Which are used to show the output shown in the below figure and also used to derive the finger location in particular note area.

4.2 Background Subtraction

A common technique for spotting moving objects in films taken by stationary cameras is background removal. The idea behind the method is to identify moving objects by comparing the current frame to a reference frame, often known as the "background picture" or "background model."[18] The backdrop picture must, at a minimum, be a depiction of the scene without any moving objects and must be changed often to accommodate changing geometry settings and brightness circumstances. The definition of the term "background subtraction"[18] has been expanded by more sophisticated models.

To figure out the location of the finger on the different notes is important which can depicts the exact location of the notes that are played sequentially. First, we had used background subtraction technique to subtract the background of the image and

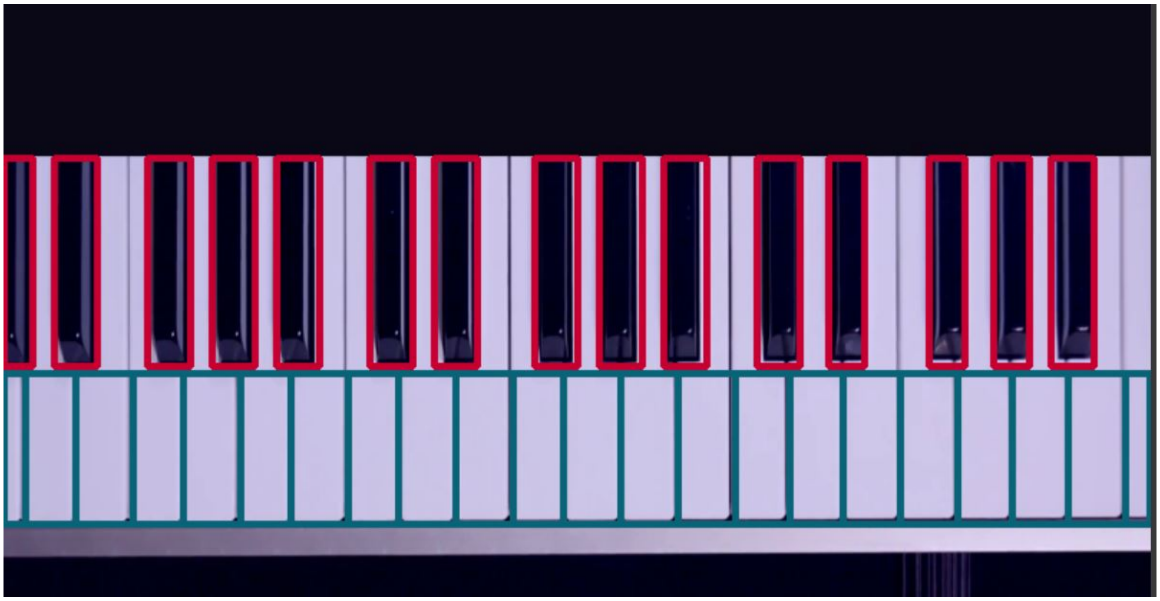


Figure 4.1: Boundary of White and Black notes.

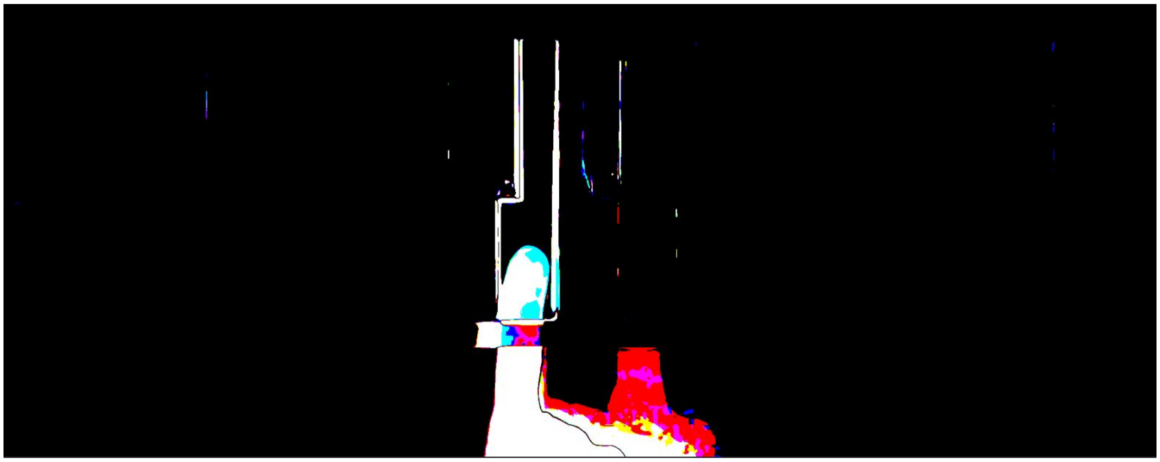


Figure 4.2: The finger detection using Background subtraction

detect the location of the object. As shown in figure 4.2, when the motion happens in the particular frame, it is compared with ideal frame and perform subtraction, it is presented in the white color. However, we can not figure out the location of the finger tip when there is a shadow in the image, and we can not find out which note is played black or white when the motion happens in both the note.

4.3 Motion Detection Technique

In video based analytic projects, motion detection is often seen. The issue may be resolved by contrasting the image's changeable and static elements, which allows one to tell the difference between the backdrop and moving items. The fundamental approach is based on matrix subtraction and averaging. The algorithm can be described as below:

- Resizing the image in order to remove unnecessary details
- reduce noise by using Gaussian blur
- perform frame subtraction of current and ideal frame
- remove pale fragments and reset them
- find the contours of moving object

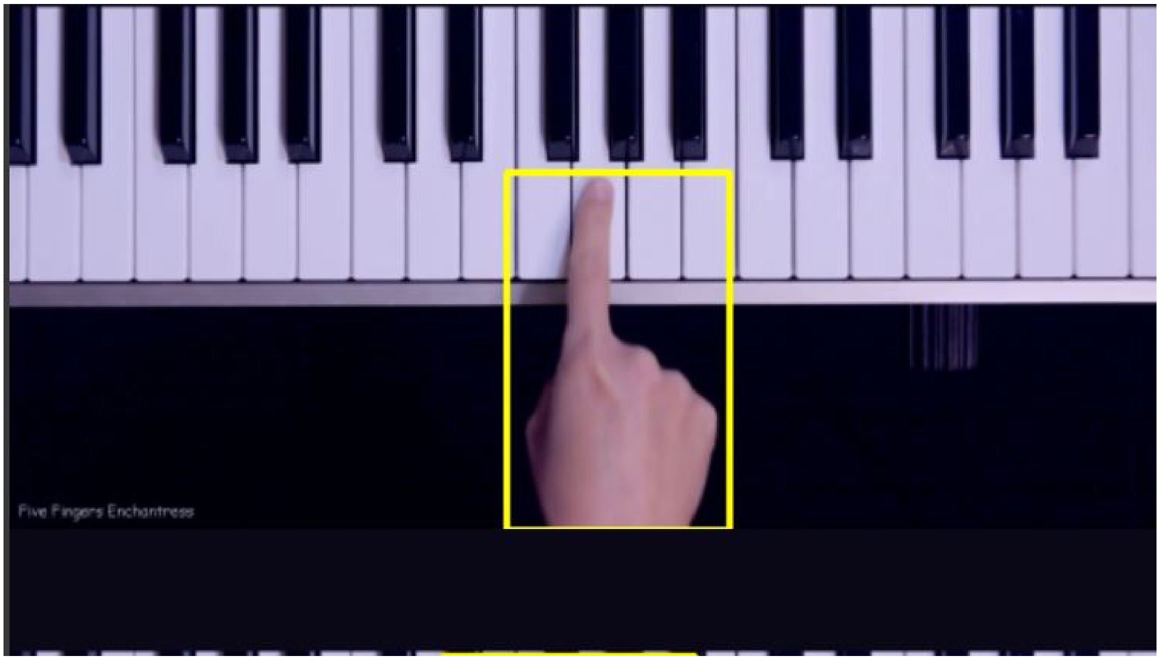


Figure 4.3: The finger detection using motion detection

Figure 4.3 describes the n^{th} frame where the motion of the hand is detected and bounded by the yellow box. It is accurate to find out the motion in the area though it is not possible to find the location of finder tip which is on the one of the notes.

Chapter 5

Multi Modal based Music Note Detection

5.1 Media Pipe - Hand Tracking Method

Hand tracking is an important component in AR/VR for providing a natural approach for engagement and communication, and it has been a hot area of study in the industry. For many years, vision-based hand position estimation has been explored. Previously published work necessitated the use of specialised hardware, such as depth sensors. Other methods are too heavy to run in real time on common mobile devices, and are thus restricted to systems with strong CPUs. Fan Zhang[19] and colleagues[10] describe a revolutionary approach that does not require any additional hardware and runs in real-time on mobile devices in their study[19].

as shown in the figure 5.1, the figure tip is always followed using the Machine Learning model provided by the mediapipe library. So, next thing we need to do is to

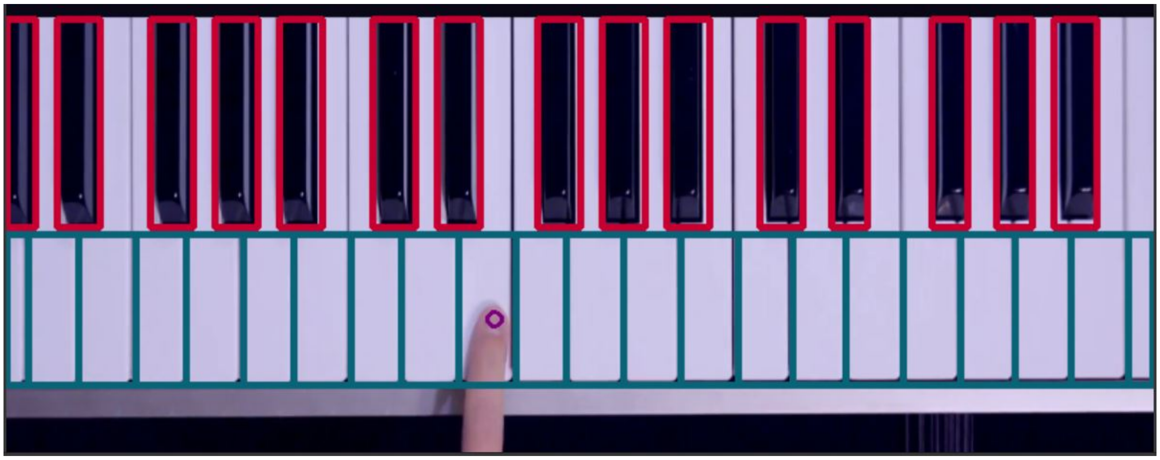


Figure 5.1: The finger detection using media pipe

find out the frame when the node is played that is done by the audio detection and recognition.

5.2 Kalman Filter

R.E. Kalman's[20] classic work providing a recursive solution to the discrete-data linear filtering issue was published in 1960. The Kalman filter has been the topic of much research and application since that time, thanks in great part to improvements in digital computing, notably in the area of autonomous or aided navigation.

The Kalman filter is a collection of mathematical equations that allows the least-squares technique to be solved in a computationally efficient (recursive) manner. The filter is extremely strong in numerous ways: it can estimate past, present, and even future states, even when the exact nature of the modelled system is unknown.

The goal of this study is to give a basic understanding of the discrete Kalman filter. This introduction comprises a derivation, description, and explanation of the basic discrete Kalman filter, as well as a relatively simple (tangible) example using real numbers, as well as a derivation, description, and discussion of the extended Kalman filter. & results[20].

The Kalman filter's main notion is that it produces a forward projection state or forecasts the next state by using past information of the state.

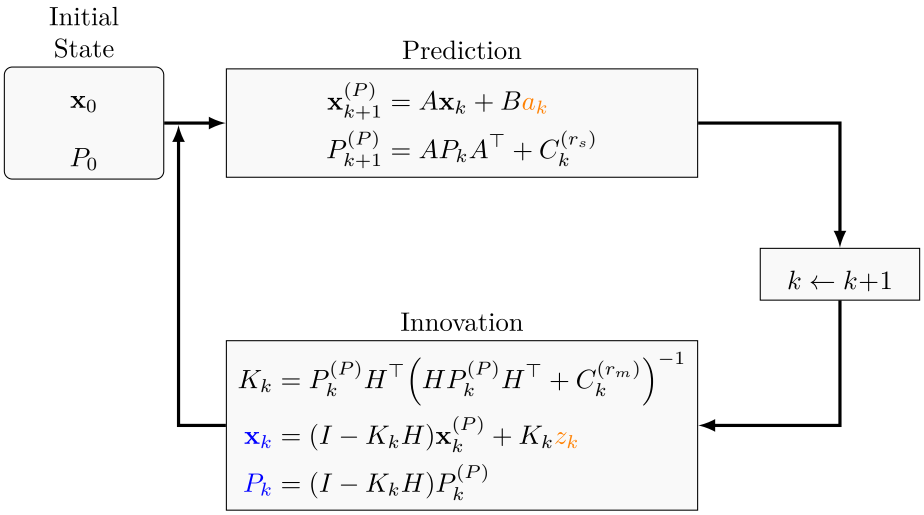

Initial State
$\mathbf{x}_0$
P_0
Prediction
$\mathbf{x}_{k+1}^{(P)} = A\mathbf{x}_k + Ba_k$
$P_{k+1}^{(P)} = AP_kA^\top + C_k^{(r_s)}$
$k \leftarrow k{+}1$
Innovation
$K_k = P_k^{(P)}H^\top\left(HP_k^{(P)}H^\top + C_k^{(r_m)}\right)^{-1}$
$\mathbf{x}_k = (I - K_kH)\mathbf{x}_k^{(P)} + K_kz_k$
$P_k = (I - K_kH)P_k^{(P)}$

Figure 5.2: Flowchart of Kalman Filter Algorithm

The Kalman Filter is used to estimate the state of a system at time $k+1$ using the linear stochastic difference equation, assuming that the state of a system at time

Algorithm 1 Kalman Filter

```
Input: $x_k$, $P_k$, $u_k$, $z_k$, $Q_k$, $A$, $B$, $R$
Output: $x_k$, $P_k$
Prediction:
$x_{k+1}^{(P)} = Ax_k + Bu_k$
$P_{k+1}^{(P)} = AP_k A^T + Q_k$
Update:
$k = k + 1$
$K_k = P_k^{(P)} H^T (HP_k^{(P)} H^T + R)^{-1}$
$x_k = (I - K_k H)x_k^{(P)} + K_k z_k$
$P_k = (I - K_k H)P_k^{(P)}$
return $x_k$, $P_k$
```

$k+1$ developed from the prior state at time k.

$$x_{k+1} = Ax_k + Bu_k$$

It is usually used in conjunction with the z_{k+1} measurement model, which specifies the relationship between state and measurement at the current step $k+1$. It's written like this [21]:

$$z_{k+1} = Hx_{k+1}$$

where:

- The state transition matrix A, a $n \times n$ matrix, connects the previous time step k to the current state $k+1$ [21].

- The control input matrix B, which is a $n \times 1$ matrix, is applied to the optional control input U_k [21].

- The transformation matrix H, which is a $m \times n$ matrix, translates the state into the measurement domain [21].

- The process noise vector with the covariance Q and the measurement noise vector with the covariance R are represented by w_k and v_k , respectively. They have a normal probability distribution and are statistically independent Gaussian noise.

Algorithm 2 Multimodal kalman filter for piano note detection using audio and video

```
Input: $N$ number of video frames and corresponding audio signal
Output: Corresponding piano notes for each frame
Procedure:
 Step1: Slice audio signal inot $N$ number of equal parts for each video frame from the framerate information
 Step2: Initialize state information $X_1$ and $z_1$(fingertip position and frequency information from the video and audio data of the first frame), $P_1$, $u_1$, $Q_1$, $A$, $B$, and $R$
 for videoframe, $k = 1$ to $N$:
     ($x_{k+1}$, $P_{k+1}$ )= KALMAN FILTER ($x_k$, $P_k$, $u_k$, $z_k$, $Q_k$, $A$, $B$, $R$)[Algorithm 1]
     Estimate the piano notes from $x_k$using nearest neighbour algorithm
```

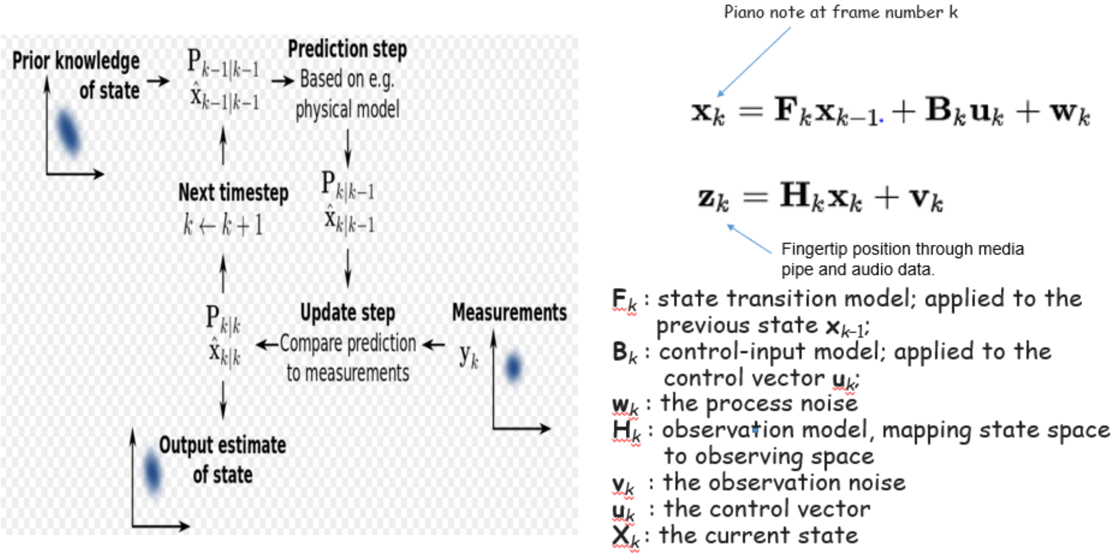

Prior knowledge of state
$P_{k-1|k-1}$
$\hat{x}_{k-1|k-1}$
Prediction step
Based on e.g. physical model
Next timestep
$k \leftarrow k+1$
$P_{k|k-1}$
$\hat{x}_{k|k-1}$
$P_{k|k}$
$\hat{x}_{k|k}$
Update step
Compare prediction to measurements
Measurements
y_k
Output estimate of state
Piano note at frame number k
$\mathbf{x}_k = \mathbf{F}_k\mathbf{x}_{k-1} + \mathbf{B}_k\mathbf{u}_k + \mathbf{w}_k$
$\mathbf{z}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k$
Fingertip position through media pipe and audio data.
$\mathbf{F}_k$: state transition model; applied to the previous state $\mathbf{x}_{k-1}$;
$\mathbf{B}_k$: control-input model; applied to the control vector $\mathbf{u}_k$;
$\mathbf{w}_k$: the process noise
$\mathbf{H}_k$: observation model, mapping state space to observing space
$\mathbf{v}_k$: the observation noise
$\mathbf{u}_k$: the control vector
$\mathbf{X}_k$: the current state

Figure 5.3: Multimodal kalman filter

The kalman filter in the proposed solution works as shown in the figure 5.3. The X_k is the resultant note for the n^{th} frame and Z_k is the fingertip position of the n^{th} frame which is derived with mediapipe and audio data. Figure 5.2 and Algorithm 1 demonstrate the flow chart and algorithm for kalman filter. Algorithm 2 illustrates multimodal kalman filter algorithm for piano note detection from audio and video data.

Algorithm 3 PARTICLE FILTER

```
Input: $\chi_k$, $u_{k+1}$, $z_{k+1}$
Output: $\chi_{k+1}$
Initialization: $\chi_{k+1} = \emptyset$
Prediction:
for $m = 1$ to $M$ do
    sample $x_{k+1}^{[m]} \sim p(x_{k+1}|u_{k+1}, x_k^{[m]})$ // Predict the state of the particle using a motion model
    $w_{k+1}^{[m]} = p(z_{k+1}|x_{k+1}^{[m]})$ // Update the weights of the particles based on current observation
    $\chi_{k+1} = \chi_{k+1} + < x_{k+1}^{[m]}, w_{k+1}^{[m]} >$ // Estimate the current state
endfor
update:
for $m = 1$ to $M$ do
    draw $i$ with probability $\propto w_{k+1}^{[i]}$ // Resample particles based on their weights
    add $x_{k+1}^{[i]}$ to $\chi_{k+1}$
endfor
return $\chi_{k+1}$
```

5.3 Particle Filter

A group of Monte Carlo techniques known as particle filters are used to address filtering issues that arise in Bayesian statistical inference and signal processing. When partial observations are taken and random disturbances are present in both the sensors and the dynamical system, the filtering issue entails estimating the internal states in dynamical systems. Given the erratic and incomplete data, the goal is to determine the posterior distributions of the states of a Markov process.In 1996, Del Moral[22] originally used the phrase "particle filters" to demonstrate mean-field interacting particle techniques that had been employed in fluid mechanics since the early 1960s. Liu and Chen introduced the phrase "Sequential Monte Carlo" in 1998[23].

Given noisy and/or incomplete data, particle filtering employs a set of particles (also known as samples) to represent the distribution of a stochastic process. The starting state and noise distributions can have any shape, and the state-space model can be nonlinear if necessary. By not making any assumptions about the state-space model or the state distributions, particle filter methods offer a tried-and-true methodology[22], [23] for producing samples from the desired distribution. However, when used on very high-dimensional systems, these strategies do not work effectively.[24]

When given observation variables, a particle filter seeks to estimate the posterior density of the state variables. In a hidden Markov Model, which has both hidden and visible variables, the particle filter is created. The hidden variables (state-process) and the observable variables (observation process) are connected by a known functional form. Similar to that, a probabilistic understanding of the dynamical system representing the development of the state variables exists. Utilizing observation measurement process, a general-purpose particle filter calculates the posterior distribution of the hidden states. Regarding a state-space like the one stated below:

$$\begin{array}{ccccccccccc} X_0 & \rightarrow & X_1 & \rightarrow & X_2 & \rightarrow & X_3 & \rightarrow & \cdots & & \text{signal} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \cdots & & \\ Y_0 & & Y_1 & & Y_2 & & Y_3 & & \cdots & & \text{observation} \end{array}$$

The above equation is observed from the book written by pierre Del Moral[25] on stochastic processes.

We can define filtering problem is, to find out the value of state X_k sequentially, from the given values of the process $Y_0, \cdots, Y_k$, at any step k.

The Markov process of $X_0, X_1, \cdots$ on $\mathbb{R}^{d_x}$ (for some $d_x \geqslant 1$) that changes in accordance with the density of transition probabilities $p(x_k|x_{k-1})$. with an initial probability density$p(x_0)p(x_0)$, this model is frequently expressed synthetically as

$$X_{k+1}|X_k = x_{k+1} \sim p(x_{k+1}|x_k)$$

In some state space, the observations Y_0, Y_1 uses values on $\mathbb{R}^{d_y}$ (for some $d_y \geqslant 1$) which are not dependent ,conditionally when $X_0, X_1, \cdots$ are known. To be precise, each Y_{k+1} is only depending on X_{k+1}. In addition, we have assumption that conditional derivation for the Y_{k+1} given $X_{k+1} = x_{k+1}$ are completely continuous ,thus, we have

$$Y_{k+1}|X_{k+1} = y_k \sim p(y_{k+1}|x_{k+1})$$

where:

- X_{k+1} is a current state/ signal.
- Y_{k+1} is observation.

Algorithm 4 Multimodal particle filter for piano note detection using audio and video

```
Input: $N$ number of video frames and corresponding audio signal
Output: Corresponding piano notes for each frame
Procedure:
 Step1: Slice audio signal inot $N$ number of equal parts for each video frame from the framerate information
 Step2: Initialize sensor information $z_{k+1}$, $\chi_k =< x_k, w_k >$(fingertip position and frequency information from the video and audio data of the first frame), and $u_{k+1}$
 for videoframe, $k = 1$ to $N$:
     $\chi_{k+1}$ = PARTICLE FILTER ($\chi_k$, $u_{k+1}$, $z_{k+1}$)[Algorithm 3]
     Estimate the piano notes from $\chi_{k+1}$ using nearest neighbour algorithm
```

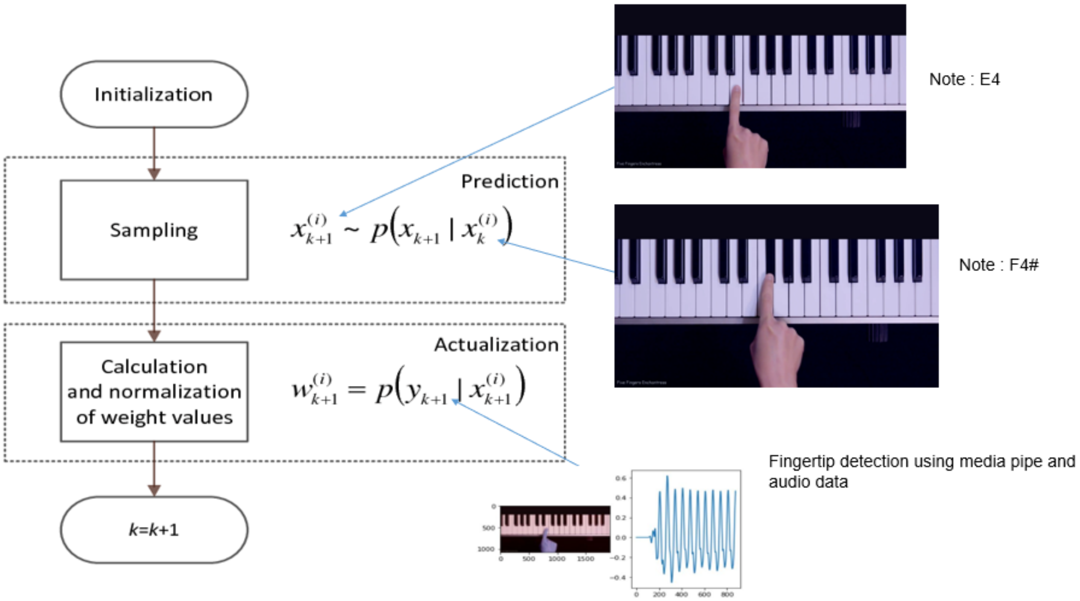

Initialization
Prediction
Sampling
$x_{k+1}^{(i)} \sim p(x_{k+1} \mid x_k^{(i)})$
Actualization
Calculation and normalization of weight values
$w_{k+1}^{(i)} = p(y_{k+1} \mid x_{k+1}^{(i)})$
k=k+1
Note : E4
Note : F4#
Fingertip detection using media pipe and audio data

Figure 5.4: Multimodal particle filter

In the context of proposed solution, the particle filter will take the n^{th} frame as input and derive the output for $n^{th}+1$ frame. As shown in the 5.4, x_k is the input frame and x_{k+1} is the output frame. During the actualization phase, the finger tip location and the audio data is given as input and the result comes in term of note. Algorithm 3 demonstrate the flow chart and algorithm for particle filter. Algorithm 4 illustrates multimodal particle filter algorithm for piano note detection from audio and video data.

Chapter 6

Results and Discussions

In figure 6.1, The graph includes 4 lines which represent the position of x coordinate from four methods manual, mediapipe, kalman filter and particle filter.

- Real position: yellow track Using mediapipe Machine learning model, we have derived finger tip predicted locations which are considered highly accurate.
- Measurements : Blue track The real position of the finger tip is derived manually which is most accurate.
- predicted : green track The predicted location is the output of the particle filter which is better than uni modal audio approach.
- predicted : Red track The predicted location is the output of the kalman filter which is pretty good than all of three modals.

Table 6.1: Modals and Accuracy

Modality	Accuracy
Unimodal(Audio signal)	73.99 %
Multi-modal(audio,video) Kalman filter	**87.81%**
Multimodal(audio,video) Particle filter	78.06%

As can be seen in the table of the modal and accuracy, the accuracy of multi modal based approach is better than uni modal approach. Moreover, kalman filter gives better results to predict the notes from audio and video data.

In the figure 6.1 and 6.2, the yellow and blue track are rarely visible due to the overlapping of the resultant track.

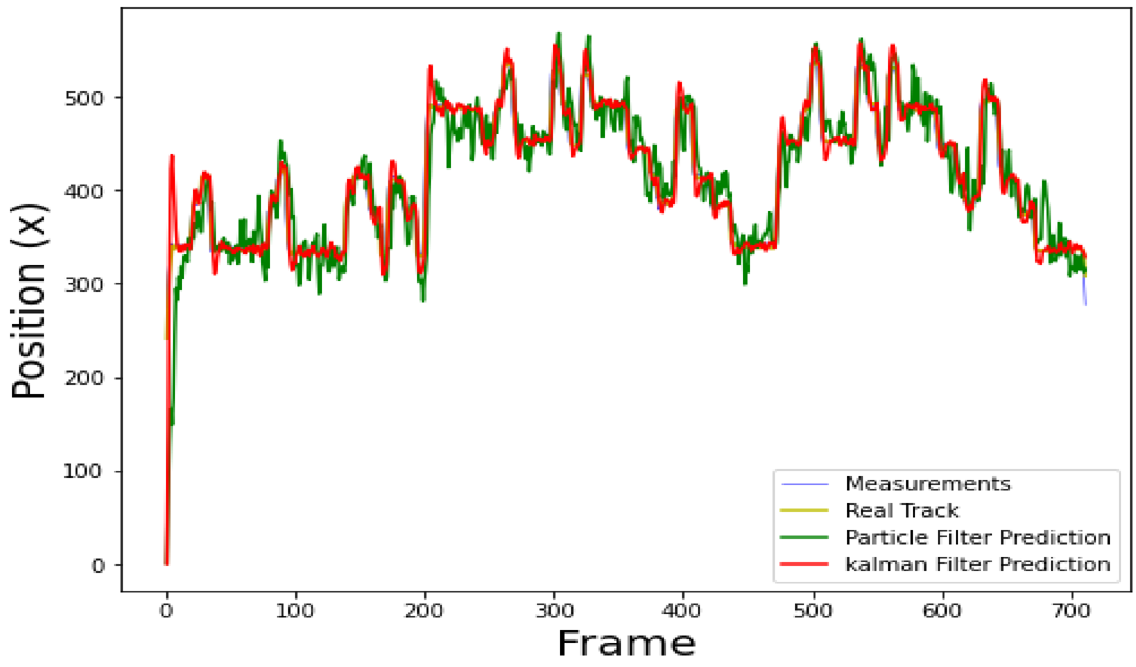


Figure 6.1: X position of each modal for each frame

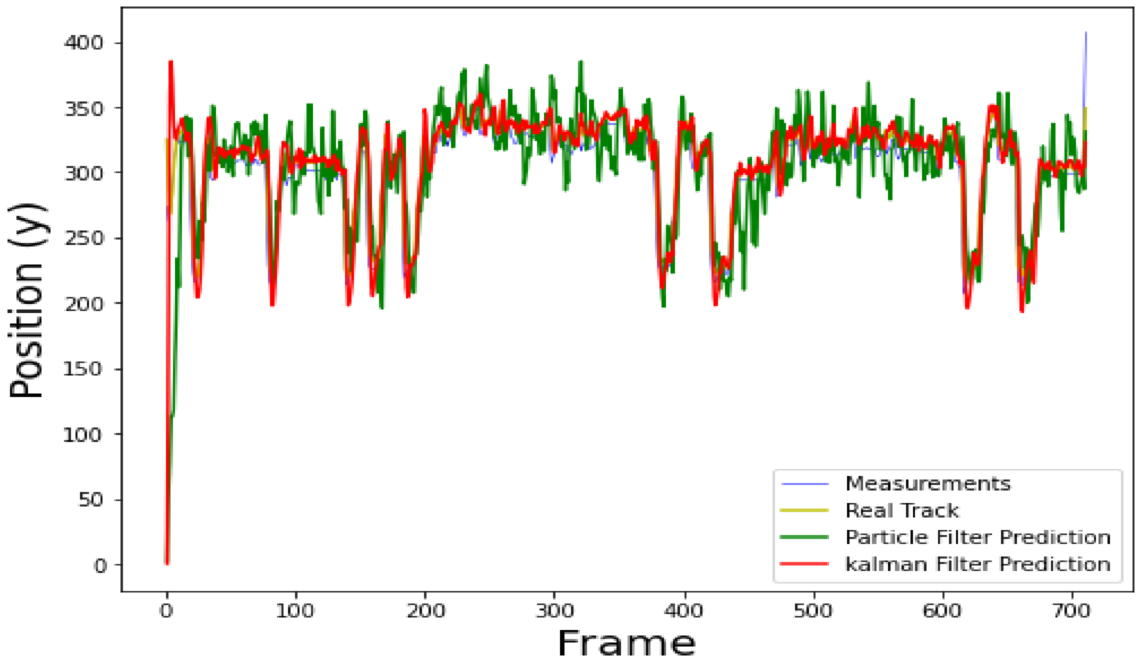


Figure 6.2: Y position of each modal for each frame

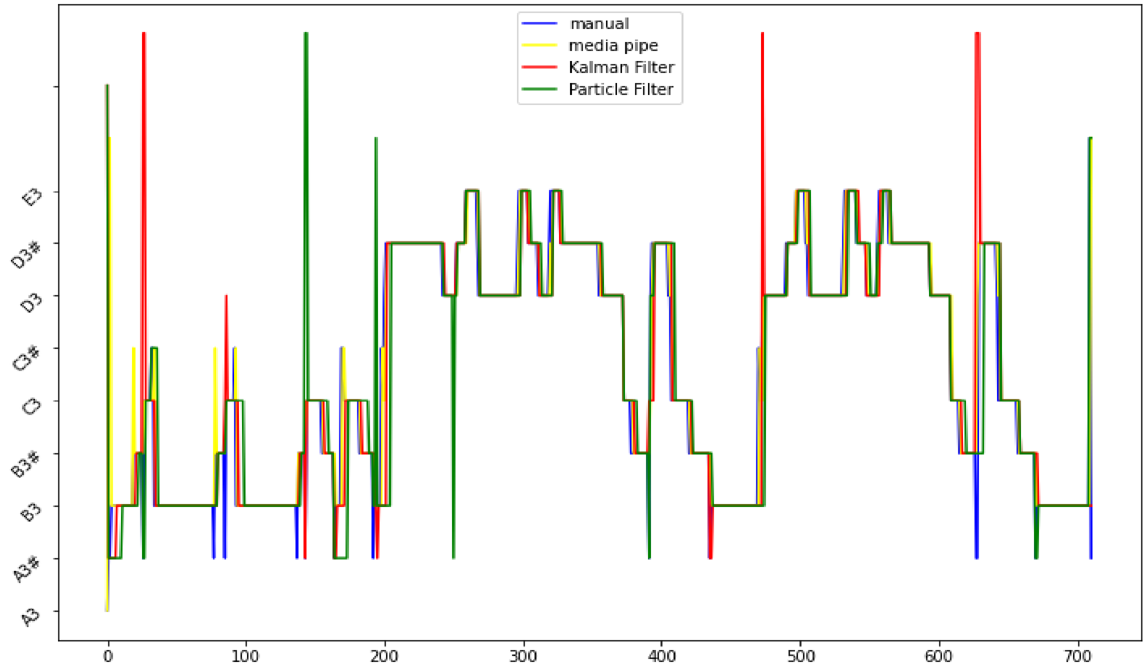


Figure 6.3: Output of the methods

In the illustrated figure 6.3, The four tracks blue, yellow, red and green depicts the output or predicted notes by manual method, media pipe method, kalman filter and particle filter, respectively, which are used through out the research. The figure has total of 711 frames and 9 notes which are frequently played. Here, the track which is above the last note(*E3*) is the demonstration of error or undetected note by the modal.

Chapter 7

Conclusion and Future Works

Music transcription and image processing are the major area of interest for multifarious rationales such as voice recognition, instrumental analysis and more. This research focuses on implementing multi modal approach to address the open problem of piano music note detection. The proposed solution has proven the usefulness of video and audio signal processing for advancement of results and accuracy. This solution has potential to be useful in other application such as speech recognition, various music sound transcription. In the solution, 851 frames are captured and processed with image based filters known as kalman filter and particle filter along with audio signals. Thee audio based analysis and note recognition for piano music consists of various known errors and faulty results that are affected due to unwanted sound known as "Noise". It is nearly impossible to eliminate the noise from the audio and minimization of the noise after certain level can alter the results so given solution helps to overcome this problem where we can derive notes not just from audio but from video too. This ensures that the detected note with audio signal processing is right. The kalman filter is commonly used for signal processing but it can be useful in the audio and video based processing and prediction which is demonstrated in the proposed solution. The particle filter which is also used for signal processing and prediction purposes is used to predict the next step with input of audio and video,however, as per the result, it is not accurate as the kalman filter. Multimodal solution outperforms conventional unimodal piano note detection algorithm. Currently, most of the available solutions have unimodal based approach where they consider only audio as input and it has accuracy of 73.99% and highly sensitive to noise. Proposed solution with multimodal based approach with particle filter gives the 78.06% accurate result which slightly higher than unimodal, however, kalman filter gives enormously accurate result which is 87.91% and almost 15% more accurate than unimodal based apporach.

This algorithm is very generic in nature and image processing and signal community can get benefited by using it for speech recognition. In future, we would like to tailor this algorithm for other signal processing applications. In the proposed solution, we have used one finger based piano music tutorial video to figure out the solution and accuracy. In future, we are going to expand this solution for multiple finger based solution where we can detect the position of all the fingers and can derive more accurate resultant notes. In the current solution we have used only piano music as input but we are aiming to use mix audio input with multiple instruments and we can separate all the audio frequencies and recognize and predict the instrument and next notes.

Bibliography

[1] D. S. A. AL-BAIYAT, *Study of kalman filter.*

[2] H. Subramanian, P. Rao, and S. Roy, *Audio signal classification*, 2004.

[3] D. Gerhard, *Audio signal classification: History and current techniques.* Citeseer, 2003.

[4] G. Velikic, E. L. Titlebaum, and M. F. Bocko, *Musical note segmentation employing combined time and frequency analyses*, IEEE, 2004.

[5] G. Peeters, *Chroma-based estimation of musical key from audio-signal analysis.* 2006.

[6] C. Yeh, A. Roebel, and X. Rodet, "Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 6, pp. 1116–1126, 2009.

[7] W. Chai and B. Vercoe, *Detection of key change in classical piano music.* 2005.

[8] ghiurutan et al., *Machine learning based pitch detection.* [Online]. Available: https://www.academia.edu/7560072/Musical_Pitch_Detection_Using_Machine_Learning_Algorithms.

[9] M. Marolt, A. Kavcic, and M. Privosnik, *Neural networks for note onset detection in piano music*, Citeseer, 2002.

[10] B. Bagby, D. Gray, R. Hughes, Z. Langford, and R. Stonner, *Simplifying sign language detection for smart home devices using google mediapipe*, 2021.

[11] X. Li, K. Wang, W. Wang, and Y. Li, *A multiple object tracking method using kalman filter*, 2010. DOI: 10.1109/ICINFA.2010.5512258.

[12] Y. Boers and J. Driessen, *Particle filter based detection for tracking*, 2001. DOI: 10.1109/ACC.2001.945669.

[13] G. Hu and D. Wang, *Auditory segmentation based on onset and offset analysis*, 2007. DOI: 10.1109/TASL.2006.881700.

[14] [Online]. Available: https://pypi.org/project/pydub/.

[15] [Online]. Available: https://en.wikipedia.org/wiki/Constant-Q_transform.

[16] [Online]. Available: https://ccrma.stanford.edu/~jos/sasp/Continuous_Wavelet_Transform.html.

[17] idrumtune.com. [Online]. Available: https://www.idrumtune.com/ultimate-guide-to-musical-frequencies/.

[18] M. Piccardi, *Background subtraction techniques: A review*, 2004. DOI: 10.1109/ICSMC.2004.1400815.

[19] F. Zhang, *Hand tracking*, 2020. [Online]. Available: https://arxiv.org/pdf/2006.10214.pdf.

[20] G. Welch, G. Bishop, *et al.*, *An introduction to the kalman filter*, 1995. [Online]. Available: https://perso.crans.org/club-krobot/doc/kalman.pd.

[21] [Online]. Available: https://en.wikipedia.org/wiki/Kalman_filter.

[22] [Online]. Available: https://people.bordeaux.inria.fr/pierre.delmoral/delmoral96nonlinear.pdf.

[23] [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10473765.

[24] [Online]. Available: https://en.wikipedia.org/wiki/Particle_filter#cite_note-2.

[25] P. Moral and S. Penev, *Stochastic Processes: From Applications to Theory*, ser. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2017, ISBN: 9781498701860. [Online]. Available: https://books.google.ca/books?id=5C%5C_vDwAAQBAJ.