ABSTRACT

This paper describes our work on the retrieval of polyphonic notes from a musical performance. Current state of the art transcription and score following systems extract features from the audio signal which are used to estimate what has been played in order to transcribe or find a position within a musical score. We propose that by extracting features from a video signal and fusing them with the audio features, the robustness of a system can be improved. We offer a framework which can be used to integrate these two modalities, and review preliminary work.

1. INTRODUCTION

Information retrieval from polyphonic music is hard. A look at the MIREX 2007 evaluation results for multi F0 estimation and note tracking shows the highest ranked systems achieves an accuracy of 0.605 [12]. Anything that can help improve on the robustness of current state of the art to enable truly usable systems would be a welcome addition.

Our visual senses are important to us when music is performed. Musicians communicate with each-other visually sometimes in very subtle ways in order to synchronise their actions. Our visual senses also help us to engage with a piece of music; it is more than just the sound that is affecting the listener when attending a live performance.

Until now, video has been largely ignored in music information retrieval systems but there is a wealth of information available within it. Transcription systems may benefit from improved accuracy through video processing although clearly there are situations where it would not be practical to use video, e.g., transcribing prerecorded collections. Our interest is in transcription for performance analysis and score following for automatic accompaniment. Unobtrusive videoing of a performance which can improve the outcome would be helpful in these cases.

For this reason, one of the goals of this work is to build a system that uses a camera pointed at the performer rather than mounted on the instrument. Another aim is to build a system that can function using consumer level equipment.

Features from the video signal can be extracted using computer vision techniques. By tracking the instrument and performer in the scene information can be retrieved, such as which notes are accessible at a particular moment or whether a new note may have been played. This information is of a fuzzy quality but may well improve on an audio-only system since it will have independence from it.

This work focuses on note recognition of single polyphonic instrument performances with string instruments. Initial experiments have been conducted using guitar although we aim to experiment with other string instruments.

This paper is structured as follows: section 2 discusses work related to ours; section 3 explains the framework we are using to develop a multimodal system; section 4 explains the approach we are using and discusses the choices made for this system; section 5 explains where we are with this system and what is planned for the future.

2. RELATED WORK

The audio component of this system is based on [3]. Contains Non-negative Matrix Factorisation (NMF) with a set of note templates learnt offline for the particular instrument being analysed. This choice of technique is discussed in section 4.1.

Combining audio and video input has had promising results in speech recognition systems [2], but to our knowledge, there is very little work in the music information retrieval field. Some work has been done using video to retrieve fingering positions for a performance [7, 1].

Burns and Wanderley [1] developed a system to retrieve the finger positions of a guitarist while they perform using a camera mounted on the neck of the guitar. This system uses image differencing to identify and track the hand of the guitarist. The finger positions were estimated using a Hough transform to find circular (finger tip) parts of the hand contour. Finger positions were then output from the system every time the hand movement was judged to be stationary. Because of the position of the camera and the camera resolution, only the first five frets could be analysed. As the hand moved further down the neck self occlusion became more significant.
3. FRAMEWORK

Our multimodal framework is shown in figure 1.

The main components are:

1. a vision system
2. an onset detector
3. an F0 component
4. a note estimator

We have chosen to use NMF to extract fundamentals from the input signal. The output of the F0 component is a frame by frame set of weights representing the contribution from each of the templates. This needs to be further processed to calculate which of these templates are significantly strong so as to represent an actual note. A further function of the note estimator is to process the frame by frame list of component notes and generate an actual transcription of the audio system including start and end time for each note.

Approaches to data fusion tend to assume the various modes contain the same information and are synchronised. For example in [8] one approach is to create a joint feature vector from multiple sources and train a system to recognise features using pattern matching techniques. [15] differentiates between data fusion and multimodal integration where the two modalities complement each other and overcome each others weaknesses but do not necessarily fuse together.

In our case, features extracted from audio and video streams are not always synchronised. The finger positions at a particular moment do not define which notes are sounding in the audio stream because the musician may have played some notes and then moved their hand.

We shall choose to inspect the video stream at the instant of the detected note onset to account for this. The note templates in the NMF contain all accessible notes from the video sequence at the last note onset and the notes that may still be sounding from previous onsets.

Evaluation is key in the development of a system of this kind, and attempts at standardisation of evaluation methods and test data sets are beginning to be used thanks to the MIREX community [5]. The difficulty in evaluating a multimodal system is it cannot be tested alongside other systems using existing audio-only test data. We shall collect audio and video inputs and evaluate an audio only system against our prototype.

4. METHOD

4.1. Extracting musical features

Several approaches to extracting notes from a performance are available. [4] explains many current techniques. For our system we need a technique that can be used alongside video. Of the techniques used in current state of the art systems, NMF was chosen for initial experiments because we are hoping to improve on audio-only results using prior knowledge that certain notes are not currently accessible (e.g. if the fingers are detected at a particular position on the neck of a guitar then many notes can be eliminated from the estimate of notes in the audio signal). When using NMF to factor an input matrix as described in [3] the templates represent each possible note for a given instrument and the result is per-frame weights given to each template in order to reconstruct an approximation of the input frame. This allows us to discard templates which represent notes that are not currently possible given the video output.

4.2. Extracting musical features from a video stream

Visual features of interest depend on the particular interface that an instrument has. For string instruments, both hands are of interest and their relative positions to the instrument as well as the gestures they make. With a right-handed guitarist, for example, the left hand position on the guitar neck constrains the pitches that can be played so it is useful to extract this from the video. It may also be useful to extract information about the hand movement because it will generally be stationary when a note is played, but this may be confused by anticipatory placements of action fingers [1].

The right hand fingers hit the strings in order to make a sound so it is useful to know where they are in relation to each string. Contact with the strings by either hand may generate sounds or change the pitch of the current sound, for example, hammer-ons and pull-offs. It may even be useful to look at the whole body of the musician as an indication of tempo and phrasing within a piece of music.
Our preliminary work achieves:

- performer location
- guitar neck location
- fretting hand/finger location

### 4.2.1. Performer location

The musician is located in the scene using skin colour and movement detection techniques. Skin detection is achieved using Gaussian Mixture Models (GMM) as described in [10] together with the dataset as published in that paper.

Foreground detection using Stauffer Grimson adaptive mixture models [14] is used in order to eliminate the background where other skin coloured regions may cause difficulties to the skin colour detection process.

Using this combination of techniques it is possible with most frames to divide the skin pixels into three regions representing the face and hands of the musician. When a frame failed to make a clear segmentation of these features it was discarded.

![Input frame](image1)

Figure 2. Segmentation of face and hands using foreground detection and skin colour detection

### 4.2.2. Guitar neck location

The guitar neck is assumed to be somewhere around and between the hands of the guitarist. Assuming the neck is the longest pair of parallel lines in this region it can be found using computer vision line detection techniques.

The frets are located on the normalised guitar neck image again using line detection. These fret candidates are used to estimate the bridge position. Then frets are compared with their neighbours. Frets that are not the correct distance from a neighbour are rejected. Correctness is assessed using the "seventeen rule" used by Luthiers. The distance between two adjacent frets is the distance between the furthest fret and the bridge divided by approximately seventeen. Equation 1 shows how to more accurately calculate this distance, where $F$ is the distance from a fret to the bridge.

$$\frac{F_n}{F_{n+1}} = \frac{\sqrt{2}}{17}$$ (1)

Of the remaining frets, an improved prediction of the location of the bridge can be made.

### 4.2.3. Tracking

After an initialisation phase a tracker is used to improve the reliability of locating the guitar neck.

Tracking is done by means of a particle filter [9]. The parameters used for the tracker are the x and y co-ordinates of top left corner the neck and the neck rotation in the frame. The length and width of the neck are already estimated via a voting mechanism from the first $n$ frames using the neck detection described previously. For simplicity, the particle transitions are just a random walk.

### 4.2.4. Fretting hand/finger location

The skin recognition GMM data set was adequate for locating the hands in the scene but was not good enough to get an accurate contour of the hand on the neck of the guitar for later processing. The original data set is used to bootstrap the process but then Expectation Maximisation (EM) is used to update the GMM parameters [11], using the algorithm explained in [13]. To achieve this, an area of pixels that are known to be skin is needed. Once the neck is located the region is then cropped and searched for skin pixels. This search looks for an RGB cluster that is spatially localised.

Running the K-means algorithm on the RGB colour values of the neck showed promising results. A binary image is formed from the cluster. This can be used to find the contour of the hand (after some cleaning-up) and subsequently the pixels within the contour are used to update the GMM via EM.

We are working on identifying fingertips from the detected hand pixels to reduce the list of accessible notes.
This is complicated by finger self occlusion which makes it hard to find all the fingers of the hand, and barre chords because when a barre is used we need more than just the fingertip position.

Initial work will not look at hand model based approaches. As discussed in [7] model based approaches are not well suited to this situation because the hand is facing away from the camera rather than trying to communicate to the camera. We are also working towards a real-time system so a computationally simple approach is desirable.

With the current video resolution (720x576) the mean neck width is around 20 pixels which is insufficient to extract accurate positions of the six strings on the neck. Higher resolution video may be required to improve the accuracy of finger position estimation.

4.2.5. System output

The system outputs the set of all notes that could possibly be accessible at this instant based on the fretting hand position on the guitar neck (see Figure 4). We are aware of a number of improvements that we might introduce to this simple method.

5. RESULTS AND CONCLUSION

A prototype vision system has been built which can identify approximate finger positions on a guitar neck and successfully constrain a list of accessible notes used by an audio system. A separate audio system is also being built based on NMF but using only the note templates that will be output from the vision system (although the two systems are not yet integrated). These preliminary results can be improved upon in future systems.

We are confident that NMF is the correct approach for this work. It has been used in state of the art audio systems and we have shown how it can be adapted to incorporate extra constraints. We believe there is a wealth of information available in the video stream that we can make use of in our future work.

6. REFERENCES