

POLYPHONIC NOTE TRACKING USING MULTIMODAL RETRIEVAL OF MUSICAL EVENTS

Garry Queded
School of Computing
University of Leeds
garryq@comp.leeds.ac.uk

Roger Boyle
School of Computing
University of Leeds
roger@comp.leeds.ac.uk

Kia Ng
School of Music
University of Leeds
kia@icsrim.org.uk

ABSTRACT

This paper addresses the problem of retrieval of musical events from a 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 using these features in conjunction with the audio features, the robustness of a system can be improved upon. We offer a framework which can improve on current state of the art by making use of these two modalities.

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 (where accuracy is measured between 0 and 1 according to the method described in [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 to the set of tools available.

Our visual senses are important to us when music is performed. Musicians visually communicate with each other 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. This suggests that may contain useful information

Until now, video has been largely ignored in music information retrieval systems but there is a wealth of information available within a video of the performance. 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 music collections. Our interest is in transcription of performances for performance analysis and score following for automatic accompaniment. Unobtrusive videoing of a performance which can improve the outcome would be helpful. 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 any

kind of mounted camera on the instrument. Another aim is to build a system that can function using consumer level equipment. This will maximise the systems usefulness in the real world.

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.

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

This paper is structured as follows: section 2 discusses work related to ours; section 6.1 outlines the current state of the art for audio feature extraction and discusses the choices made for this system; section 6.2 discusses techniques used to extract features from the video signal; section 3 discusses issues of how to combine the two modalities to improve on audio only feature extraction; section 5 discusses issues arising when trying to evaluate such a system; section 8 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 [4]. Cont uses 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 6.1.

Combining audio and video input has had promising results in speech recognition systems [3]. 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].

[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 also used difference image 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. Fin-

ger 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.

[6] describes a system which transcribes drum sequences using a combination of features from audio and video to improve on a previous, audio only system. In this work the camera is placed unobtrusively. Our work has similar aims to that of Gillet in that we aim to make a multimodal system that is unobtrusive for the musician and improves on single mode feature extraction. The actual techniques used in Gillet’s work are not used here.

3. DATA FUSION

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 compliment each other and overcome each others weaknesses but don’t 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. This makes data fusion less straight-forward.

One possible solution to this with our approach is to only use the video output at the moment a note onset occurs.

4. FRAMEWORK

Our multimodal framework is shown in figure 1.

The main components are:

1. a vision system
2. an onset detector
3. an f_0 component
4. a note estimator

We have chosen to use NMF to extract fundamentals from the input signal so the output of the f_0 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.

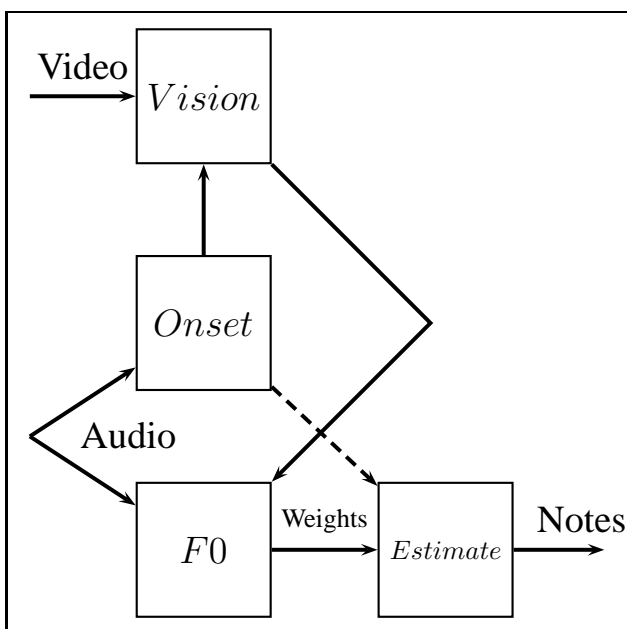


Figure 1. Framework for multimodal polyphonic note estimation

The note templates in the NMF contains all accessible notes from the video sequence at the last note onset and the notes that may still be sounding from previous onsets.

5. EVALUATION

Standard methods of evaluating transcription systems and score followers including test sets are beginning to be used thanks to the MIREX community [5]. The difficulty with a new multimodal system is it cannot be tested alongside other systems using existing audio-only test data. One thing that can be evaluated is the improvement achieved using a multimodal system compared to the same system running with only audio input.

Evaluation will therefore be done as follows. Extract features from a pre-recorded performance that has been manually labelled and aligned with a score. Firstly, evaluate accuracy using audio only input. Secondly, use audio and video inputs. The assessment of correctness of retrieval is based on the MIREX evaluation of multi f_0 estimation and note tracking.

6. METHOD

This section discusses the approach we have taken when building this system and explains why certain decisions were made.

6.1. Extracting musical features

Several approaches to extracting notes from a performance are available. [2] explains many of the techniques and is a great resource. For our system we need a technique that can be used alongside the video system. We also require

a system that is simple to implement so we can work on the computer vision and data fusion components of our project. Non-negative Matrix Factorisation was chosen for initial experiments because we are hoping to improve on audio only results using our 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 [4] the templates represent each possible note for a given instrument and the results of the NMF is per-frame weights given to each template in order to reconstruct an approximation of the input frame. This allows us to throw away templates which represent notes that are not currently possible given the video output.

6.2. Extracting musical features from a video stream

Visual features of interest depend on the particular interface that an instrument has. For string instruments, both the musicians 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 made by the guitar so it is useful to extract this from the video. It may also be useful to extract information about the movement of this hand because it will generally be stationary when a note is played. This is not always useful due to anticipatory placements of action fingers as discussed in [1].

The right hand fingers hit the strings in order to make a sound so it is useful to know where the fingers of the right hand 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 a hammer-ons where the left hand finger sharply presses a string causing a note to sound. It even may be useful to look at the whole body of the musician as an indication of tempo and phrasing within a piece of music.

Our current vision system can be divided into the following components:

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

6.2.1. Performer Location

The musician is located in the scene using skin colour detection and movement detection techniques. Skin detection is achieved using the Mixture Of Gaussian approach described in [10] also using the dataset published in that paper.

Foreground Detection using Stauffer Grimson adaptive mixture models [14] 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.

6.2.2. 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 the following technique:

1. the image is cropped around the region of interest
2. an edge image is created of the region
3. the image is repeatedly rotated and projected onto the Y-axis
4. the projection with the two largest peaks is assumed to be the correct rotation of the neck.

The frets are located on the neck using the normalised guitar neck image using a vertical projection similar to the one described above. The fret candidates are then 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.

$$F_n/F_n = \sqrt[17]{2} \quad (1)$$

Of the remaining frets, a new prediction of the location of the bridge can be made.

6.2.3. Tracking

The current system uses a tracker to improve the reliability of locating of the guitar neck and then processes this neck to find a hand contour but an initial boot strapping phase is needed before the tracker can start.

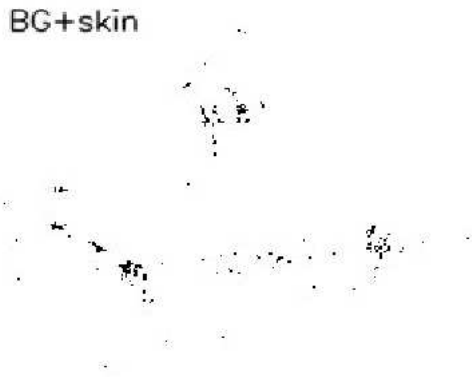
Tracking is done by means of a particle filter [9]. The parameters used for the tracker are the x and y co-ordinates of top left corner the neck and the neck rotation in the frame. The length and width of the neck are already estimated via a voting mechanism from the first n frames using the neck detection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

6.2.4. Fretting Hand/Finger Location

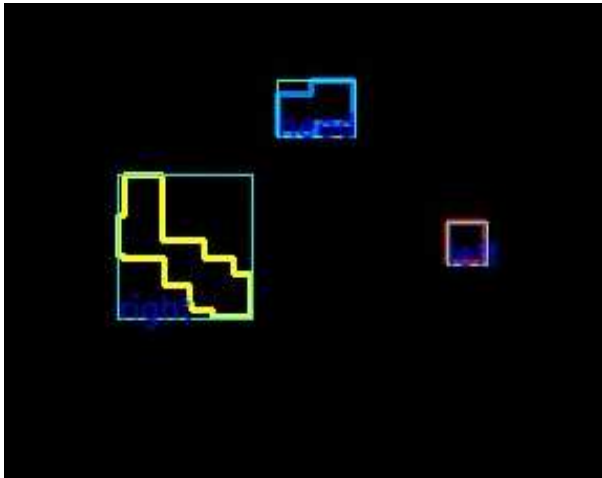
The skin recognition Gaussian Mixture model (GMM) using the data set published in [10] was adequate for locating the hands in the scene but was not good enough to get an accurate contour of the hand on the neck of the guitar for



(a) Input frame



(b) Skin detection: black pixels show areas where skin has been detected



(c) Segmentation of detected pixels into 3 regions: head, left-hand, right-hand

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



Figure 3. Running K-means on the guitar neck RGB pixels using 3 means



Figure 4. Results of image processing. The system outputs the set of all frets partially covered by the hand are output so that all possible pitches in the audio signal can be calculated.

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 uses the fact that a guitar neck is very uniform in its features. The finger board is one colour, the frets another, and possibly the markers another colour. These are all spread across the length of the neck whereas a hand would only take up a short section of the length of the neck.

Running the K-means algorithm on the RGB colour values of the neck showed promising results. Three means gave the best results in tests and the hand can clearly be seen in Figure 3. To decide which index contains the skin pixels the variance of x position is taken for the pixels in each index. The index with the lowest variance is taken to be the mean that represents mostly hand pixels because all the other indexes should represent features which are spread across the whole neck.

A binary image is then made from these pixels. This binary image 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.

6.2.5. System output

The system then outputs the set of all notes that could possibly be accessible at this point in time based on the fretting hand position on the guitar neck (see Figure 4. This is currently a very crude estimation but there are many ways in which it can be improved.

6.2.6. Finger detection

I am currently working on identifying fingers from the detected hand pixels. This is complicated by the fact that often the finger tips are not visible (self occlusion) and sometimes one finger will either be holding down more

than one string (barre chords) or may not be touching the strings at all. With the current video resolution the string positions must be estimated. This is because the neck width averages around 20 pixels. This may require a higher resolution camera if the results are not accurate enough. Initial work will not look at hand model based approaches. This is for a number of reasons: 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, the visual data is only required to compliment the audio data so we do not have to pin-point the finger positions on the neck, we just need to reduce the number of possible pitches to make note event recognition easier.

7. RESULTS

This research is still in its early stages and results will be published in due course. So far a prototype vision system has been built which can identify approximate finger positions on a guitar neck and output a list of possible notes.

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).

8. CONCLUSION

Already this project is showing that significant benefits can be achieved from a multimodal approach to music information retrieval. NMF is well suited to this type of system as it is easily adapted to varying the templates for each frame. There is a wealth of information in the video stream that we have not even begun to process yet so there is a lot of possibilities for future work.

9. REFERENCES

- [1] Anne-Marie Burns and Marcelo M. Wanderley. Visual methods for the retrieval of guitarist fingering. In *NIME '06: Proceedings of the 2006 conference on New interfaces for musical expression*, pages 196–199, Paris, France, France, 2006. IRCAM — Centre Pompidou.
- [2] Alain De Cheveigne. *Multiple F0 Estimation*, chapter 2, pages 45–79. Hoboken, NJ: Wiley inter-science; Chichester: John Wiley [distributor], 2006.
- [3] CC Chibelushi, F. Deravi, and JSD Mason. A review of speech-based bimodal recognition. *Multimedia, IEEE Transactions on*, 4(1):23–37, 2002.
- [4] A. Cont. Realtime Multiple Pitch Observation using Sparse Non-negative Constraints. 2006.
- [5] J.S. Downie, K. West, A. Ehmann, and E. Vincent. The 2005 music information retrieval evaluation exchange (mirex 2005): Preliminary overview. *Proc. 6th International Symposium on Music Information Retrieval ISMIR*, pages 320–323, 2005.
- [6] Olivier Gillet and Gael Richard. Automatic transcription of drum sequences using audiovisual features. In *IEEE ICASSP*. Springer, 2005. <http://ieeexplore.ieee.org/iel5/9711/30652/01415682.pdf?arnumber=1415682>.
- [7] D. O. Gorodnichy and A. Yogeswaran. Detection and tracking of pianist hands and fingers. In *Computer and Robot Vision, 2006. The 3rd Canadian Conference on*, pages 63–63, June 2006. <https://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-48492.pdf>.
- [8] DL Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, 1997.
- [9] Michael Isard and Andrew Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.
- [10] Rehg James M Jones Michael J. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.
- [11] S.J. McKenna, Y. Raja, and S. Gong. Tracking colour objects using adaptive mixture models. *Image and Vision Computing*, 17(3):225–231, 1999.
- [12] G.E. Poliner and D.P.W. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.
- [13] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. Thomson Learning Vocational, 1998.
- [14] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of the IEEE Computer Science Conference on Computer Vision and Pattern Recognition (CVPR-99)*, pages 246–252, Los Alamitos, June FebruaryMarch–FebruaryMay 1999. IEEE.
- [15] A. Tanaka and R.B. Knapp. Multimodal interaction in music using the Electromyogram and relative position sensing. *Proceedings of the 2002 conference on New interfaces for musical expression*, pages 1–6, 2002.