COMBINED AUDIO AND VIDEO ANALYSIS FOR GUITAR CHORD IDENTIFICATION

Alex Hrybyk and Youngmoo Kim
Electrical & Computer Engineering, Drexel University
{ahrybyk, ykim}@drexel.edu

ABSTRACT

This paper presents a multi-modal approach to automatically identifying guitar chords using audio and video of the performer. Chord identification is typically performed by analyzing the audio, using a chroma based feature to extract pitch class information, then identifying the chord with the appropriate label. Even if this method proves perfectly accurate, stringed instruments add extra ambiguity as a single chord or melody may be played in different positions on the fretboard. Preserving this information is important, because it signifies the original fingering, and implied "easiest" way to perform the selection. This chord identification system combines analysis of audio to determine the general *chord scale* (i.e. A major, G minor), and video of the guitarist to determine *chord voicing* (i.e. open, barred, inversion), to accurately identify the guitar chord.

1. INTRODUCTION

The ability of an instrument to produce multiple notes simultaneously, or chords, is a crucial element of that instrument's musical versatility. When trying to automatically identify chords, stringed instruments, such as the guitar, add extra difficulty to the problem, because the same note, chord, or melody can be played at different positions on the fretboard. Figure 1 depicts a musical passage in staff notation, followed by three representations in tablature form (the horizontal lines represent the strings of the guitar, and number is the fret of that string). All of these tablature notations are valid transcriptions, in that they produce the correct fundamental frequencies as the staff notation when performed. However, only one of these positions may correspond to the original, perhaps easiest fingering

Guitar lessons are more accessible now than ever with the rise of streaming Internet video and live interactive lessons. The research presented in this paper has direct applications to these multimedia sources. A system which can automatically transcribe chord diagrams from audio and video lessons between student and teacher would be an invaluable tool to aid in the learning process.

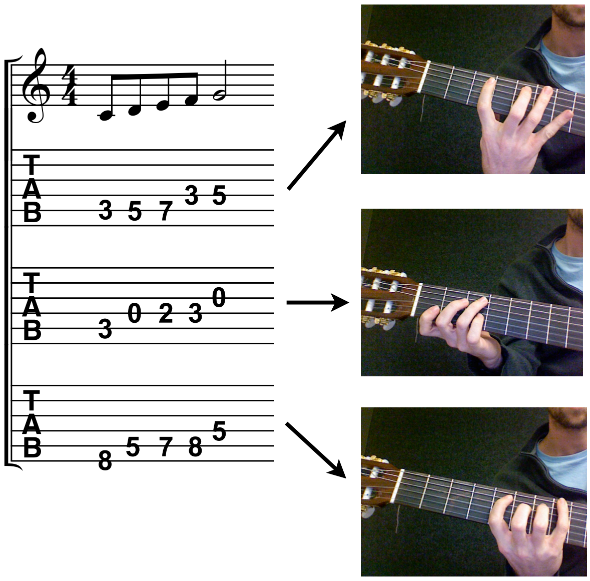

Figure 1. Three voicings of a C major scale in staff and tablature notation, shown in various positions along the guitar fretboard.

Automatic chord identification algorithms have traditionally used the chroma feature introduced by Fujishima [1]. The chroma based approach, though intuitive and easily implemented, presents many problems due to the existence of overtones in the signal. This paper avoids this problem by using a polyphonic pitch estimation method named *Specmurt Analysis* which filters out the overtones in the log-frequency spectrum to yield only a chord's fundamental frequencies [2].

Visual approaches to guitar chord and melody transcription have been attempted. Most of these methods, while accurate, are obtrusive to the guitarist; cameras must be mounted to the guitar [3], or the guitarist must wear colored fingertips to be tracked [4]. The method presented here uses brightly colored dots placed at various points along the guitar's fretboard to be tracked by the camera. These dots, which are unobtrusive to the guitarist, are used as reference points to isolate the fretboard within the image, so that principal components analysis may be used to identify the guitarist's particular voicing of that chord.

The multi-modal guitar chord identification algorithm presented in this paper is as follows: first, using Specmurt Analysis, fundamental frequency information will be retrieved and the general *chord scale* identified (i.e. G major, A# minor, etc.). Next, using video analysis, the guitarist's particular *chord voicing* (i.e. open, barred, inversion, etc.) will be identified using principal components analysis (PCA) of the guitarist's fretting hand.

2. RELATED WORK

The chromagram or pitch class profile (PCP) feature has typically been used as the starting point for most chord recognition systems. Fujishima first demonstrated that decomposing the discrete Fourier transform (DFT) of a signal into 12 pitch classes and then using template matching of various known chords produces an accurate representation of a song's chord structure [1].

The main problem with chroma is apparent when using template matching for various chords. For example, a C Major triad would have an ideal chroma vector of $[1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0]$. The existence of overtones in the signal cause the ideal 0's and 1's to fluctuate and create false chord identifications.

Modified versions of the chromagram, such as the Enhanced Pitch Class Profile by Lee have been introduced to ease the effects of overtones in the signal [5]. This method computes the chroma vector from the harmonic product spectrum rather than the DFT, suppressing higher harmonics making the chroma vector more like the ideal binary template. However, this method fails to identify the voicing of the chord, such as a first or second inversion.

Burns et al. developed a visual system for left-hand finger position tracking with respect to a string/fret grid [3]. Their method relies on the circular shape of fingertips, using a circular Hough transform on an image of the left-hand to detect fingertip locations with respect to the underlying fretboard. However, this method requires mounting a camera on the headstock of the guitar, which poses many problems: it can be obtrusive to the guitar player's natural method of playing, and also only captures information about the first five frets of the guitar.

Kerdvibulvech et al. proposed to track the fingering positions of a guitarist relative to the guitar's position in 3D space [4] . This is done by using two cameras to form a 3D model of the fretboard. Finger position was tracked using color detection of bright caps placed on each of the guitarist's fingertips. Again, this can hinder the physical capabilities and creative expression of the guitarist, which should not happen in the transcription process.

3. AUDIO ANALYSIS

When playing a single note, instruments produce natural harmonics (overtones) in addition to the note's fundamental frequency. Therefore, when playing multiple notes, the frequency spectrum of the audio appears cluttered, making detection of the fundamental frequencies (the actual notes) hard to locate. Saito *et al.* have proposed a technique called *Specmurt* analysis, which will be used to extract the notes of a guitar chord from the audio signal [2].

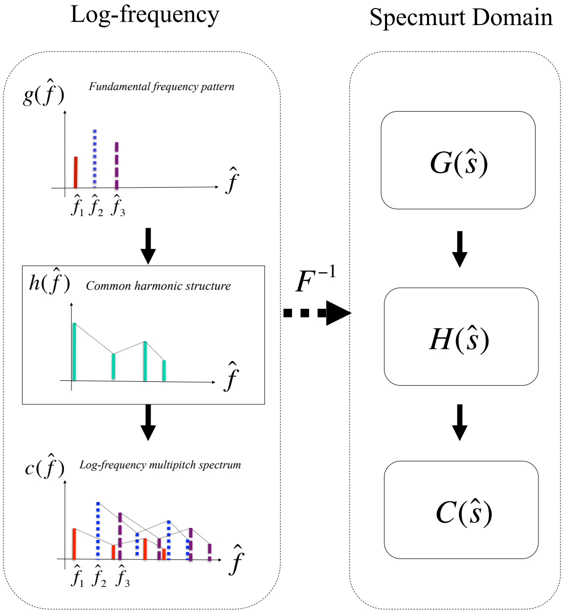
Log-frequency
Specmurt Domain
$g(\hat{f})$
Fundamental frequency pattern
$\hat{f}$
$\hat{f}_1\ \hat{f}_2\ \hat{f}_3$
$h(\hat{f})$
Common harmonic structure
$\hat{f}$
$c(\hat{f})$
Log-frequency multipitch spectrum
$\hat{f}$
$\hat{f}_1\ \hat{f}_2\ \hat{f}_3$
F^{-1}
$G(\hat{s})$
$H(\hat{s})$
$C(\hat{s})$

Figure 2. Log-spaced frequency domain $c(\hat{f})$ as a convolution of common harmonic structure $h(\hat{f})$ with fundamental frequency distribution $g(\hat{f})$.

3.1 Specmurt Analysis

Multiple fundamental frequency estimation using Specmurt analysis is performed by inverse filtering the log-scale frequency domain with a common harmonic structure of that instrument [2]. The resulting log-frequency spectrum contains only impulses located at the log-fundamental frequencies.

Harmonics of a fundamental frequency f_0 normally occur at integer multiples of the fundamental, nf_0. Furthermore, if the fundamental frequency changes by some Δf, the change in frequency of its respective higher harmonics will also be $n\Delta f$. By resampling the frequency domain to have a log-scaled axis, this allows the harmonics of a given fundamental to be consistently spaced by $\log n + \log f_0$, independent of fundamental frequency.

$$\hat{f} = \log f \tag{1}$$

3.1.1 *Common Harmonic Structure*

Using the log-scale frequency axis, we can assume that the harmonic frequencies are located at $\hat{f} + \log 2, \hat{f} + \log 3, ..., \hat{f} + \log n$. When a chord is played on an instrument, each note will presumably contain these same harmonic frequencies, beginning at different $\hat{f}$'s. Therefore, we can assume that the log-scaled multipitch spectrum, $c(\hat{f})$, is a combination of these harmonic structures, shifted and weighted differently per note. Specifically, the resulting log-scale frequency spectrum, $c(\hat{f})$, is equal to the convolution of a common harmonic structure, $h(\hat{f})$, with a fundamental frequency distribution, $g(\hat{f})$.

$$c(\hat{f}) = h(\hat{f}) * g(\hat{f}) \tag{2}$$

The harmonic structure can be written in terms of its log-frequency axis spacing, $\hat{f}_{0n}$, and its harmonic weights, W_n, where $n = 1, 2...N$ harmonics.

$$h(\hat{f}, W) = \sum_{n=1}^{N} W_n \delta(\hat{f} - \hat{f}_{0n}) \tag{3}$$

The harmonic weights will initially be a guess, which will be refined later using an iterative process to minimize the overall error of Specmurt analysis.

3.1.2 Specmurt Domain

In order to determine the desired fundamental frequency distribution, $g(\hat{f})$, one can solve (2) by deconvolving the log-spectrum with the common harmonic structure. An easier way of obtaining $g(\hat{f})$ would utilize the duality of the time/frequency-convolution/multiplication relationship (shown in Figure 2). Therefore, taking the inverse Fourier transform would yield the relationship

$$\begin{aligned} \mathcal{F}^{-1}\{c(\hat{f})\} &= \mathcal{F}^{-1}\{h(\hat{f}) * g(\hat{f})\} && (4) \\ C(\hat{s}) &= H(\hat{s})G(\hat{s}) && (5) \end{aligned}$$

where $\hat{s}$ is a temporary Specmurt domain variable. Simple algebra followed by a Fourier tranform of $G(\hat{s})$ will yield the resulting fundamental frequency spectrum.

$$G(\hat{s}) = \frac{C(\hat{s})}{H(\hat{s})} \tag{6}$$

$$\mathcal{F}\{G(\hat{s})\} = g(\hat{f}) \tag{7}$$

The squared error after performing Specmurt analysis can be defined as

$$E(W_n) = \int_{-\infty}^{+\infty} \left\{ c(\hat{f}) - h(\hat{f}, W_n) * g(\hat{f}) \right\}^2 d\hat{f} \tag{8}$$

Minimizing the error of Specmurt is done by refining the harmonic weights, W_n, of the harmonic structure. This is done by setting the error's N partial derivatives $\frac{\partial E}{\partial W_n} = 0,\ n = 1...N$, and solving the system of equations for W_n.

The original Specmurt formulation assumed that the first weight, $W_1 = 1$, of the normalized common harmonic structure. After experimentation with various guitar signals, the higher harmonics were sometimes of larger magnitude than the fundamental frequency. By allowing the first harmonic's magnitude to vary, the algorithm was able to better identify fundamental frequencies.

4. VIDEO ANALYSIS

In order to visually identify the performing guitarist's chord voicing, the guitar fretboard must first be located and isolated within the image. However, the guitar can be held in many different orientations relative to the camera, making it difficult to find the location or coordinates of the fretboard in the image plane.

The frets of a guitar are logarithmically spaced to produce the 12 tones of the western scale. The coordinates in the (x, y) plane are plotted in Figure 3, where the x_i coordinates are related by

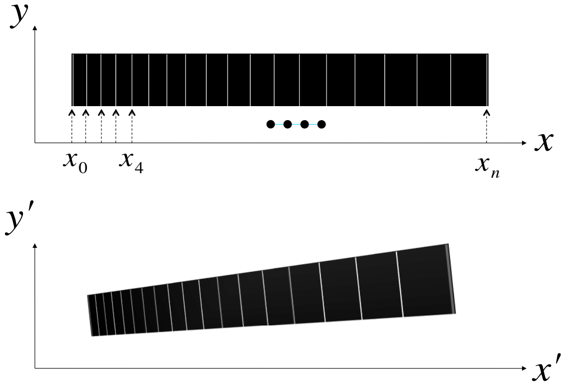

Figure 3. Ideal fretboard (top) with logarithmic x spacing of n frets, and arbitrary neck width in y direction, and seen image (bottom) with warped spacing.

$$x_i = \sum_{k=0}^{i} x_0 \times 2^{\frac{k}{12}} \tag{9}$$

4.1 Homography

Homography is the process of applying a projective linear transformation to a scene (a 2D image or 3D space), to describe how perceived positions of observed objects change when the point of view of the observer (a camera) changes. Homography will be used to determine the correct perspective transformation, i.e. rectify or warp the original image to fit the ideal fretboard spacing in Figure 3. This will make it easy to isolate the fretboard in the image for analysis. The general homography matrix equation

$$w\mathbf{p}' = \mathbf{H}\mathbf{p} \tag{10}$$

states that points in the image, $\mathbf{p}'$ can be expressed as a warping of ideal points $\mathbf{p}$ with a homography matrix $\mathbf{H}$, including a scale factor w. The homography matrix is a transformation matrix between the two images, based on which a one-to-one relationship between the features points $\mathbf{p}'$ and $\mathbf{p}$ [6]. Specifically, the points will have two dimensions, x and y, and will be expressed in terms of a 3x3 homography matrix with elements h_{ij}.

$$w \begin{bmatrix} x_i' \\ y_i' \\ 1 \end{bmatrix} \approx \begin{bmatrix} h_{00} & h_{10} & h_{20} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \tag{11}$$

x_i are determined from (9) and y_i are determined as an arbitrary guitar neck width (from the ideal, rectangular fretboard). The corresponding reference points (x_i', y_i') in the image now need to be established, to compute the homography matrix, $\mathbf{H}$.

4.2 Reference Point Tracking

In order to perform the homography rectification concepts in 4.1, the correct reference points in the image must be

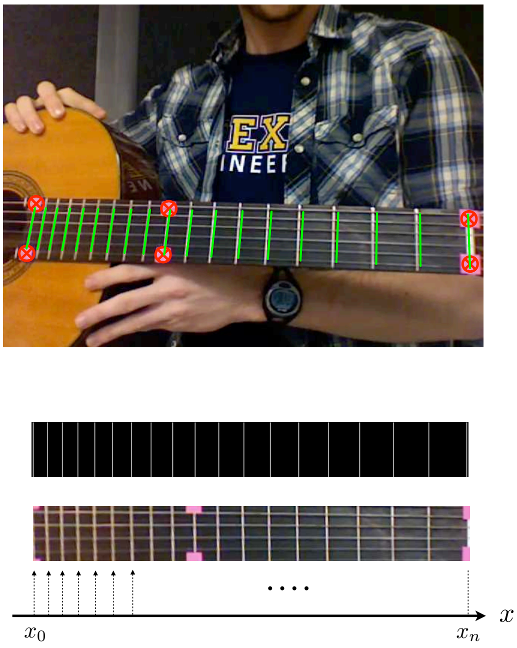

Figure 4. (top) Original image showing tracking points (in red), projected frets (in green) using the homography matrix. (bottom) Ideal fretboard, and subsection of original image after applying homography matrix to each coordinate.

determined. Attempts were made at using an iterative nonlinear error minimization method, which proved initially unsuccessful (see later section 6). Instead, distinct bright colored stickers were placed at various fret locations on the neck of the guitar. The coordinates of these points (x_i', y_i') were tracked in each frame of video using a simple color masking followed by a k-means clustering algorithm. The small stickers were placed on the neck of the guitar on either side of the metal frets, so as not to interfere with the guitarist's playing and the timbre of the instrument.

A set of (x_i, y_i) and (x_i', y_i') now exist, corresponding to the frets of the guitar. The homography matrix is determined by minimizing the mean square error of (11) using these points. Applying the inverse transformation, $\mathbf{H^{-1}}$, to the ideal grid in Figure 3 yields frets that overlay perfectly with the frets in the image (Figure 4). Applying $\mathbf{H}$ to the original image and taking the subsection of coordinates yields the rectified fretboard (Figure 4), whose fret spacings are known from (9). The rectified fretboard is now isolated and in a usable form for PCA.

4.3 Determination of Chord Style

The next goal is to determine which chord voicing, given the subset of voicings that exist for a particular chord. PCA is used to decompose the rectified fretboard in its "eigen-chord" components, and determine the correct chord voicing.

Let the training set of fretboard images be $F_1, F_2 ... F_M$ which are vectors of length LW for an image with dimensions L by W. An example training set of fretboard images is shown in Figure 5. The average image is $A = \frac{1}{M}\sum_{i=1}^{M} F_i$, and each image with subtracted mean is $\bar{F}_i = F_i - A$. PCA seeks to find the eigenvectors and eigenvalues of the covariance matrix

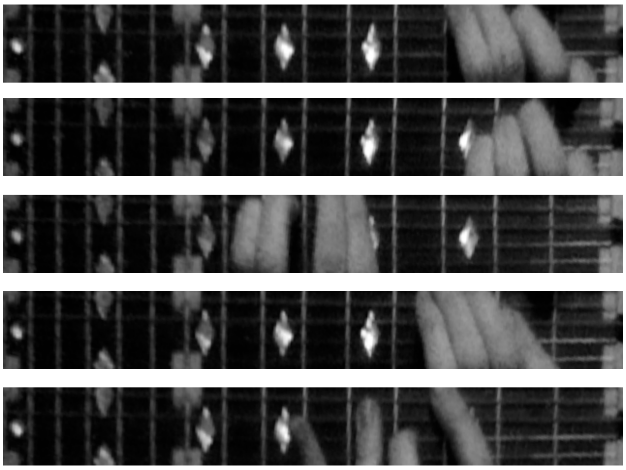

Figure 5. Example fretboard images used for training.

$$\mathbf{C} = \frac{1}{M}\sum_{i=1}^{M} \bar{F}_i \bar{F}_i^T \qquad (12)$$

$$= \mathbf{S}\mathbf{S}^T \qquad (13)$$

where $\mathbf{S} = [\bar{F}_1, \bar{F}_2 ... \bar{F}_M]$ is a set of training of images in matrix form. However, $\mathbf{C}$ is of dimension LW; the images used in this experiment are of size 80x640 pixels, and computing 51200 eigenvectors and eigenvalues is computationally intractable. Turk et. al presented a method for solving for the LW eigenvectors by first solving for the eigenvectors of an MxM matrix $\mathbf{S}^T\mathbf{S}$ [7]. The M eigenvectors v_l are used to form the eigenvectors u_l of $\mathbf{C}$.

$$u_l = \sum_{i=1}^{M} v_l \bar{F}_i \qquad l = 1...M \qquad (14)$$

A new image $\tilde{F}$ can be reduced to its eigen-chord components, c_k, using the M' eigenvectors which correspond to the larger eigenvalues of $\mathbf{S}^T\mathbf{S}$.

$$c_k = u_k(\tilde{F} - A) \qquad k = 1...M' \qquad (15)$$

5. EXPERIMENTAL RESULTS

Three guitarists were asked to perform a sequence of chords from chord diagrams. The chords were a selection drawn from eight scales (major and minor), each in three voicing-dependent positions: open (traditional open stringed), barred, and a 1st inversion, totaling 24 chords all together. The system was evaluated using various combinations of features derived from audio only, video only, and combinations thereof. All experiments were performed using leave-one-out training of audio and video when using PCA.

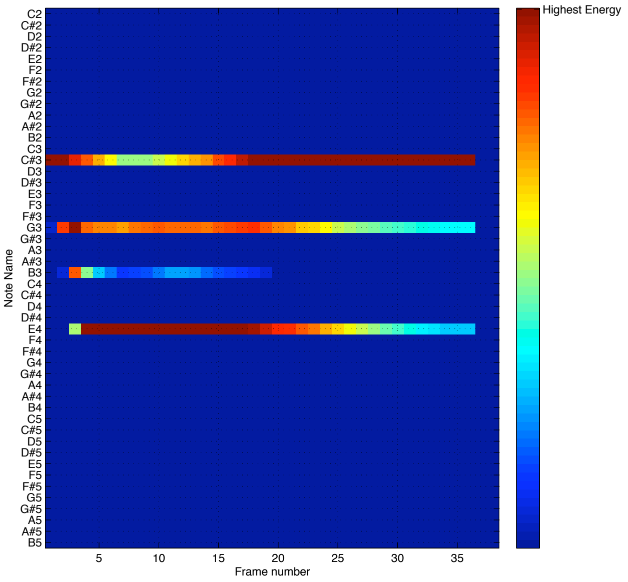

Figure 6. Specmurt piano-roll output of a C#m7♭5 jazz chord.

5.1 Audio Only

The output of Specmurt analysis is a piano-roll vector of size 48, each element corresponding to the energy of a chromatic note from C2 to B5 (4 octaves, 12 notes per octave). An example of a piano-roll vector over multiple time frames is shown in Figure 6.

Two methods were used to calculate the correctness of the chord scale and voicing using this vector. It is known what notes make up each major and minor scale. Therefore, the chord scale was evaluated by summing the energy over all octaves of the notes belonging to that chord - similar to chroma analysis. The chord scale with the highest energy was assumed to be correct, yielding an accuracy of 98.6%.

It is not deterministic, however, as to which chord voicing created a particular set of notes, or chord. For example, both the G major open chord and G major barred chord contain six notes total, all of the same fundamental frequencies, but the notes are rearranged on different strings, and hence use different fingerings. Therefore, a training set using the piano-roll energy vector was developed for each chord scale. Using PCA to identify chord voicing from the piano-roll vector shows some accuracy (62%) but is understandably low, as the difference in note energies may be very fine and inseparable for different voicings with similar notes.

	Audio only	Video only	Combined System
Scale	**98.6**	34.8	**98.6**
Voicing	62.0	**94.4**	**94.4**
Both	61.1	32.8	**93.1**

Table 1. Accuracy results for various combinations of modes of information. Combined accuracy results using Specmurt for scale identification, and video for voicing identification showing highest accuracy.

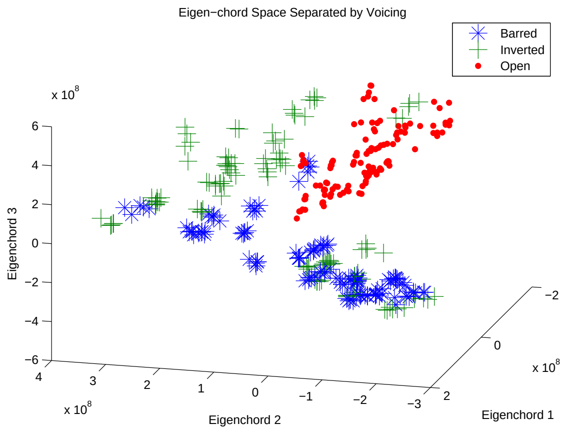

Figure 7. Three voicings from A minor, G major, and C major, after being projected into the chord-space. Various colors and symbols show the how the voicing of chords remain grouped after dimensionality reduction.

5.2 Video Only

A training set of 240 images was used to build the eigen-chord space for each test. Frames of video were then projected into the chord-space using three eigen-chords of the training set using (15), and its closest centroid was assumed to be the correct chord.

Chord scale identification using only video performed extremely poorly (34%). This is expected, as the chord scale centroids in the projected chord-space after PCA are somewhat meaningless. For a particular chord scale, many different voicings exist at various points on the fretboard, which is what we hope to separate by using PCA.

For chord voicing however, very high accuracy was achieved (94.4%). Figure 7 shows how various voicings of chords, irrespective of scale, tend to group together due to the similar hand shapes used by the guitarist.

5.3 Combined System

The system which performs the best in terms of correctly identifying the overall chord (scale and voicing) utilizes the strong points of scale and voicing identification within the audio and video results. Since Specmurt analysis yielded extremely high accuracy for determining scale, it was used as a preprocessing step to voicing identification via video.

6. FUTURE WORK

The video and audio components of this guitar chord identification system have the potential to be expanded upon.

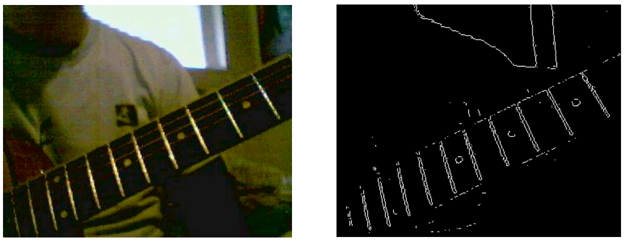

Figure 8. Guitar image (left) and edge-thresholded image (right).

6.1 Automatic Fretboard Registration

Placing colored tracking points along the neck of the guitar presents additional constraints on how the guitar fretboard can be rectified: all the tracking points must be visible in the frame of video, and nothing else in the frame may have similar color. Ideally, we would like to locate the fretboard without these points. By looking at the edge-detected image of a guitar, this produces a fairly accurate representation of where the frets are - the color of the metal frets contrasts heavily with that of the wooden neck, providing edges at frets (Figure 8).

Using the homography concept in 4.1, the points denoted as edges, $\mathbf{p}'$, should be warped using $\mathbf{H^{-1}}$ to align with the ideal fret-grid points $\mathbf{p}$. This is equivalent to minimizing an error function defined as

$$E(\mathbf{H}) = ||\mathbf{p} - \mathbf{H^{-1}p}'||^2 \tag{16}$$

$$\mathbf{H} = \underset{\mathbf{H}}{\text{argmin}}(E(\mathbf{H})) \tag{17}$$

After experimentation, the error function $E(\mathbf{H})$ is noticeably non-convex, and contains local minima in $\mathbf{H}$. The two fret-grids "align" in alternate orientations which are incorrect, but still minimize the error function. This area of research is being continued with the motive of constraining (16) and (17), such that the error function will always be convex, and converge to a global minimum when the two images are correctly aligned.

6.2 Larger Training Sets

Very high accuracy of video voicing identification (94.4%) was achieved using image data from only three guitarists. A more robust classifier of chord voicings could be created by collecting more data, to account for players who use non-traditional finger orientations for chords. With more data, the accuracy of determining chord scale from video may increase (34.8%), as scales may then form more meaningful distributions in the eigen-chord space.

6.3 Additional Chord Types

This system is very extendable to detect different chord scales besides major and minor. Detection of diminished, augmented, 7th, and other jazz chords are easily implemented with the chroma-style analysis of Specmurt's output, and refined search using the eigen-chord decomposition of the fretboard image.

6.4 Fusing Audio/Video Data

Currently the system uses Specmurt analysis to determine a chord's scale as a pre-processing step to eigen-chord decomposition of the fretboard to determine voicing. This means that any error introduced by Specmurt propagates throughout the rest of the system. Therefore it is desired to jointly estimate the scale and voicing together using audio and video features simultaneously.

7. CONCLUSION

This paper has presented an alternate approach to automatic guitar chord identification using both audio and video of the performer. The accuracy of chord identification increases from 61.1% to 93.1% when using audio for scale identification, and video for voicing. The "eigen-chord" decomposition of fretboard images proved extremely successful in distinguishing between a given chords voicings (normal, barred, inverted) if the chord scale is known (94.4%).

8. REFERENCES

[1] T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music." in *Proceedings of the International Computer Music Conference*, 1999.

[2] S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama, "Specmurt analysis of polyphonic music signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 3, pp. 639–650, February 2008.

[3] A.-M. Burns and M. M. Wanderley, "Visual methods for the retrieval of guitarist fingering," in *NIME '06: Proceedings of the 2006 conference on New interfaces for musical expression*. Paris, France, France: IRCAM — Centre Pompidou, 2006, pp. 196–199.

[4] C. Kerdvibulvech and H. Saito, "Vision-based guitarist fingering tracking using a bayesian classifier and particle filters," in *PSIVT07*, 2007, pp. 625–638.

[5] K. Lee, "Automatic chord recognition from audio using enhanced pitch class profile," in *Proceedings of the International Computer Music Conference*, 2006.

[6] X. Wang and B. Yang, "Automatic image registration based on natural characteristic points and global homography," in *Computer Science and Software Engineering, 2008 International Conference on*, vol. 5, dec. 2008, pp. 1365 –1370.

[7] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, June 1991, pp. 586 –591.