

# VISUAL PITCH CLASS PROFILE

## *A Video-Based Method for Real-Time Guitar Chord Identification*

First Author Name, Second Author Name

*Institute of Problem Solving, XYZ University, My Street, MyTown, MyCountry*  
*f.author@ips.xyz.edu, s.author@ips.xyz.edu*

Third Author Name

*Department of Computing, Main University, MySecondTown, MyCountry*  
*t.author@xy.mu.edu*

**Keywords:** Guitar Chord Identification, Chord Detection, Pitch Class Profile, Chroma Vector.

**Abstract:** We propose a video-based method for guitar chords identification which is analogous to the *state-of-the-art* audio-based method. While the method based on audio data uses the Pitch Class Profile feature and supervised Machine Learning techniques to “teach” the machine about the chord “shape”, we use as feature the approximated positions of fingertips in the guitar fretboard, captured using especial hardware. We show that visual and audio-based methods have similar classification performance, but the former outperforms the later with respect to the immunity to noise caused by strumming.

## 1 INTRODUCTION

Despite being a predominantly auditory activity, music has also a visual component which is worth noting. In the case of guitar playing, for instance, it is easier to identify chords by visual inspection than by auditive perception.

Most studies in Musician-Computer Interaction, however, have been done by the Computer Music community, and there are still few Computer Vision approaches.

In this work we propose a new method for real-time guitar chords identification using only Computer Vision techniques. The method is analogous to the *state-of-the-art* audio-based method, which uses a Supervised Machine Learning technique to train the machine with the patterns of different chords, using as training feature the Pitch Class Profile. We kept the Machine Learning part and replaced the auditory feature by a visual one, namely the approximate positions of fingertips at the guitar fingerboard.

Guitar chord detection has many applicatios. It can be used to develop interactive educative softwares, where the computer evaluates if the student is performing correctly some chord sequence. The same idea can also be applied in games, where the interface is a real guitar. Mucitians could also benefit of this kind of system by using it to control audio effects or

even some authomatic composition algorithm.

Section 2 briefly lists previous works dealing with similar problems. In Section 3 we will describe the audio-based method that inspired ours, which will be detailed in Section 4. Section 5 is the place for comparative analysis between the audio and video based mentioned techniques. Conclusions and future work ideas take place in Section 6.

## 2 PREVIOUS WORK

In (Burns and Wanderley, 2006) a camera is mounted on the guitar headstock in order to capture the first five frets. The Linear Hough Transform is used to detect strings and frets, and the Circular Hough Transform is used to locate the fingertips. The system has also a module for movement detection. The idea is to use the Hough transforms only when the hand is not moving. The purpose is to identify chords and notes sequences in real-time by detecting the fingertips positions in guitar fretboard coordinates. So the system does not use Machine Learning tools.

The work of (Kerdvibulvech and Saito, 2007) is more ambitious. They use stereo cameras and augmented reality fiducial markers to locate the guitar fingerboard in 3D, and colored markers (with differ-

ent colors) attached to the fingertips to determinate their three-dimensional position relative to the fretboard. They apply a Bayesian classifier to determine the color probabilities of finger markers (to cope with changes in illumination) and a particle filter to track such markers in 3D space. The system works in real-time.

The two mentioned works use only visual information. In (Quested et al., 2008), visual information is used to enhance the performance of an audio-based musical information retrieval method. The point is that once the fundamental frequency of the played note is known (via audio), the video information helps solving the ambiguity regarding which string was actually fingered or plucked. In real-time, the guitar is detected using edge methods, and a skin recognition technique is applied to roughly locate the position of the hand relatively to the fretboard.

The same idea is used in (Paleari et al., 2008), but their system is not designed to work in real-time. In the first video frame, the Linear Hough Transform is applied to segment the guitar from the scene, and after the image is rotated so that the guitar neck becomes horizontal, edge methods are used to locate the fretboard. After that, tracking of the fretboard points in the video is done by means of the Tomasi Lukas Kanade algorithm. The hand position is determined via skin color methods.

As can be seen, there is always a trade-off between algorithmic complexity and scene interference. The method we will describe (in section 4) also conforms with this *there-is-no-free-lunch* paradigm.

### 3 THE AUDIO-BASED METHOD

According to (Cabral, 2008), most of the audio-based chord recognition methods rely on the use of the Pitch Class Profile (PCP) audio feature along with some Supervised Machine Learning method.

Let us now define the PCP, also known as the Chroma Vector, as described by (Jehan, 2005).

At regular intervals the audio file is hann-windowed and the magnitudes of the Discrete Fourier Transform coefficients are computed. Then the 84 amplitudes corresponding to MIDI notes ranging from 24 to 107 are captured and a 12-dimensional vector is obtained by summing the amplitudes corresponding to musical notes of the same key in different octaves. The elements of this vector are normalized to the range  $[0, 1]$  to avoid taking into account differences of loudness in different windows.

Chroma vectors roughly represent the likelihood of a musical note (regardless of its octave) being

present in the audio window under analysis. Therefore for the PCP of a chord follow some pattern, which can be learned somehow by the machine.

In the seminal work of (Fujishima, 1999) the Nearest Neighbor method is used, and the machine is trained with “ideal” chroma vectors: those whose entries are 1’s in the notes of the chord and 0’s otherwise.

For the purpose of evaluating our visual method, we have implemented an audio-based chord detection algorithm similar the one introduced by Fujishima, where instead of using “ideal” chords, the machine was trained with samples from real recorded chords, and at the classification phase we have used the  $K$  Nearest Neighbor algorithm with  $K > 1$ .

## 4 PROPOSED VIDEO-BASED METHOD

For the purposes of this work, let us define the Region of Interest (ROI) in the scene of a person playing guitar as being the region including the strings, from the nut to the bridge.

To facilitate the capture process, avoiding the overhead of segmenting the ROI, we chose to work in the infrared-light range.

Figure 1 shows the equipment that supports our method. We use a infrared camera to capture the scene, which is properly illuminated with infrared light. Special markers (fiducials) are attached to the guitar in order to easily locate the instrument, and for the fingers, reflexive gloves dress the middle phalanges.

The pipeline of our chord detection method is illustrated in Figure 2. The developed software took advantage of some nice and robust algorithms implemented in OpenCV, an open source Computer Vision library (Bradski and Kaehler, 2008).

First, a threshold is applied to the input image, so that the only non-null pixels are those of the guitar and finger markers. Then, using the contour detection algorithm and contour data structure provided by OpenCV, guitar and finger markers can be separated. Note that guitar fiducials and finger markers are, respectively, contours with and without a hole. Once the positions of the four guitar fiducials are known in the image, by using their actual positions in guitar fingerboard coordinates a projective transformation (homography) can be applied in order to “immobilize” the guitar and easily extract the ROI. This homography is then applied to the northmost extreme of the finger rods, so we get the rough position of fingertips in guitar fretboard coordinates, since the distal



Figure 1: Capture hardware. On the left, an infrared camera surrounded by four infrared light sources. In the center, a hollow disk made with reflexive material. Four of them are used to locate the plane containing the ROI. On the right, middle-phalanges gloves with small rods coated so as to easily reflect light.

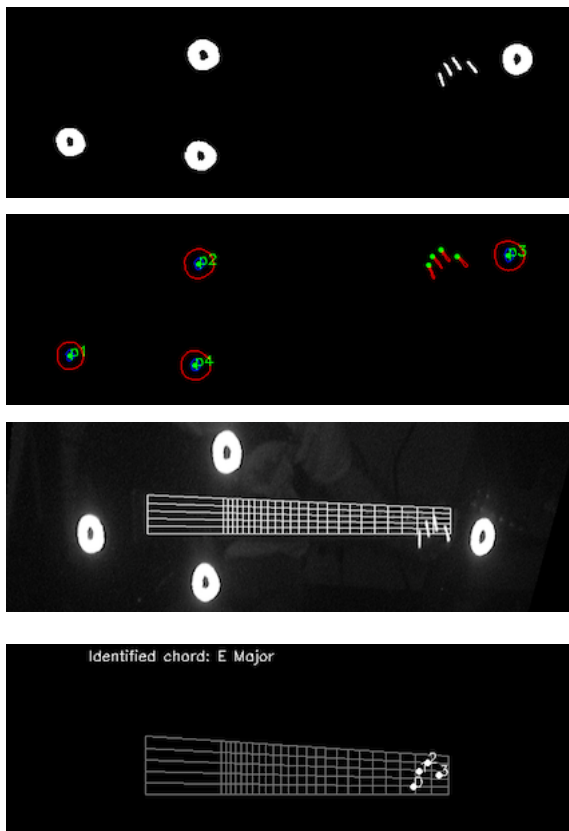


Figure 2: Chord detection pipeline, from top to bottom. (1) A threshold is applied to take only guitar and finger markers. (2) Guitar fiducials and finger rods are detected using a contour detection algorithm. (3) A projective transformation “immobilize” the guitar, regardless the movement caused by the musician. (4) The projective transform is applied to the northmost extreme of finger rods in order to roughly locate the fingertips in guitar-fretboard coordinates.

phalanges are, in general, nearly perpendicular to the fingerboard.

However, as mentioned before in fact we don’t care about the precise coordinates of fingertips. In-

stead we use a Supervised Machine Learning technique to train the machine with the guitar chords we want it to identify. The chord a musician plays is viewed by the system as an eight-dimensional vector composed by the coordinates (after projective transformation) of the four fingertips, from the little to the index finger. By analogy with the PCP, we call this eight-dimensional vector the Visual Pitch Class Profile (VPCP).

Summarizing, the proposed algorithm for real-time guitar chord detection has two phases. In the first (the training phase), the musician choose the chords that must be identified and take some samples from each one of them, where by sample we mean the eight-dimensional vector formed with the positions of the foremost extreme of the finger rods. In the identification phase, the system receives the vector corresponding to chord to be identified and classifies it using the  $K$  Nearest Neighbor algorithm.

## 5 COMPARISON AND DISCUSSION

It’s not easy to compare our chord detection algorithm with the existing video-based methods, cited in Section 2, in terms of classification performance, since there is no publicly available database for that purpose.

Despite this fact, the goal of our work is not to compete with other methods. The goal is to introduce a simple algorithm, video-based, that has some advantages regarding the analogous *state-of-the-art* audio-based method for real-time guitar chord detection. This way our method can add robustness to the algorithm based on audio data or even replace it in some situations. Moreover, the fact that both (audio and video based) algorithms are similar facilitates the application of data fusion techniques (more in Section 6).

Before talking about quantitative comparisons, let's address some theoretical aspects. Video methods, even knowledge-based, are immune to wrong tuning of the instrument. Despite not being desirable to play a wrong tuned instrument, this feature is good for beginners, that are not able to have a precisely regulated guitar. On the other hand, it can be argued that knowledge-based methods only work properly when trained by the final user itself, since the shapes of some given chord are slightly different from person to person. This is a fact, but the knowledge-based techniques using audio data also have to face with this problem, since different instruments, with different strings, produce slightly different songs for a given same chord.

Seeking quantitative comparisons, we take 100 samples from each one of the 14 major and minor chords in the keys of C, D, E, F, G, A, B, choosing just one shape per chord (in the guitar there are many realizations of the same chord). The video samples were taken by fixing a given chord and, while moving a little bit the guitar, waiting until 100 samples were saved. For the audio samples, for each chord we recorded nearly 10 seconds of a track consisting of strumming in some rhythm keeping fixed the chord. The audio data was then pre-processed in order to remove parts corresponding to strumming (where there is high noise). Then, at regular intervals of about 12 milliseconds an audio chunk of about 45 milliseconds was processed to get its Pitch Class Profile, as described in Section 3.

These audio and video samples tend to form clusters in  $\mathbb{R}^{12}$  and  $\mathbb{R}^8$ , respectively. Figure 3 provides some analysis of them. Note that in both cases the samples are placed very close to the mean of the respective cluster, but there are more outliers in the audio data.

Regarding classification performance, both methods behaved well in the tests we have conducted. The difference is that the audio-based algorithm is sensitive to the noise caused by strumming, while the video-based method don't care about it. This is illustrated in Figure 4, where the same chord sequence (played twice) was performed and analysed by the two methods. Note how more stable is the video-based method. It can also be seen that both algorithms have problems with chord transitions.

## 6 CONCLUSIONS AND FUTURE WORK

We proposed a video-based method for real-time guitar chords identification which is analogous to the

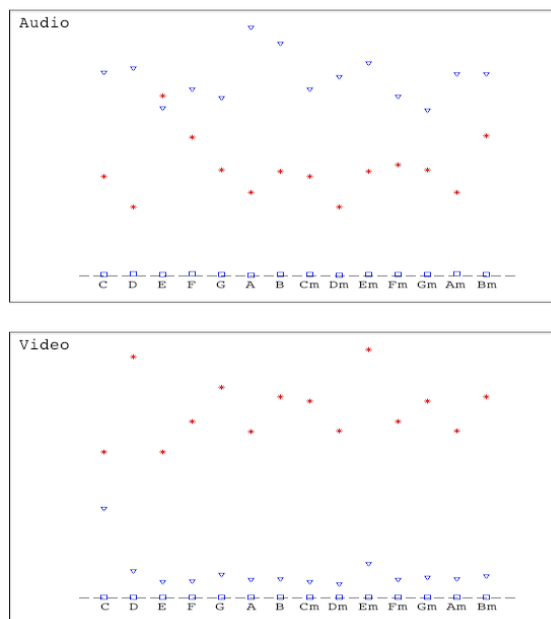


Figure 3: Analysis of the audio and video sample clusters. A square (respectively, a triangle) represent the average (respectively, the maximum) distance between the class samples and the class mean vector. The asterisk represent the distance between the cluster mean vector and the nearest cluster mean vector. This shows that the clusters of video samples are better defined relatively to those from audio samples.

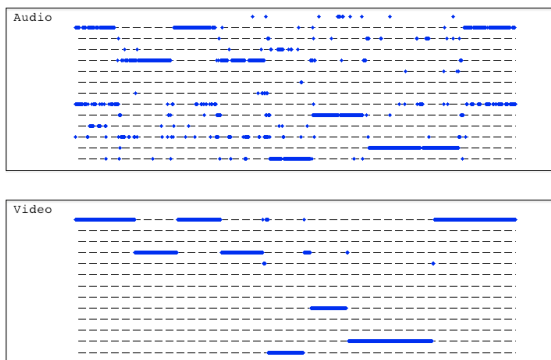


Figure 4: The same chord sequence, played twice, is analysed by the traditional audio-based algorithm and our proposed video-based method. While the former needs some extra processing to cope with the noise caused by strumming, our video based method is immune to that. However, both techniques have problems with chord transitions.

*state-of-the-art* audio-based method. Instead of using the Pitch Class Profile audio feature, our method is based on what we call the Visual Pitch Class Profile (VPCP).

We have seen that both methods have similar clas-

sification performance, but the VPCP algorithm is more stable in the sense that (1) the clusters formed at the training phase are better defined and (2) the visual method is not sensitive to the noise caused by strumming.

Given the high similarity between the classical audio-based method and our proposed video-based algorithm, a natural direction of research is to combine both classifiers using some data fusion technique.

There is also some issues of the VPCP method which have to be treated. The first is to eliminate the need of the middle-phalanges gloves. Although they are unobtrusive for the user, having a free hand is highly desirable. Also, since the rods stay out of the fretboard and we are applying a perspective (plane to plane) transformation, rotating the guitar in the direction of the neck axis causes a vertical shift in the transformed rod northmost points.

To cope with this issues we plan to use two calibrated cameras, one working in the infrared range to capture the guitar (which may have some infrared leds), and the other in the visible range to capture the fingertips using traditional hand- and contour- detection methods.

## REFERENCES

- Bradski, G. and Kaehler, A. (2008). *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly.
- Burns, A. and Wanderley, M. (2006). Visual methods for the retrieval of guitarist fingering. In *Conference in New Interfaces for Musical Expression*.
- Cabral, J. (2008). *Harmonisation Automatique en Temps Reel*. PhD thesis, Universite Pierre et Marie Curie.
- Fujishima, T. (1999). Real-time chord recognition of musical sound: A system using common lisp music. In *International Computer Music Conference*.
- Jehan, T. (2005). *Creating Music by Listening*. PhD thesis, Massachusetts Institute of Technology.
- Kerdvibulvech, C. and Saito, H. (2007). Vision-based guitarist fingering tracking using a bayesian classifier and particle filters. *Advances in Image and Video Technology*.
- Paleari, M., Huet, B., Schutz, A., and Slock, D. (2008). A multimodal approach to music transcription. In *15th International Conference on Image Processing*.
- Quested, G., Boyle, R., and Ng, K. (2008). Polyphonic note tracking using multimodal retrieval of musical events. In *International Computer Music Conference*.