Estimation of Guitar Fingering and Plucking Controls based on Multimodal Analysis of Motion, Audio and Musical Score.

Alfonso Perez-Carrillo[1,3], Josep-Lluis Arcos[2], and Marcelo Wanderley[3]

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
alfonso.perez@upf.edu,
WWW home page: http://www.dtic.upf.edu/~aperez/
[2] IIIA-CSIC, Barcelona, Spain
[3] IDMIL, McGill University, Montreal, Canada

Abstract. This work presents a method for the extraction of instrumental controls during guitar performances. The method is based on the analysis of multimodal data consisting of a combination of motion capture, audio analysis and musical score. High speed video cameras based on marker identification are used to track the position of finger bones and articulations and audio is recorded with a transducer measuring vibration on the guitar body. The extracted parameters are divided into left hand controls, i.e. fingering (which string and fret is pressed with a left hand finger) and right hand controls, i.e. the plucked string, the plucking finger and the characteristics of the pluck (position, velocity and angles with respect to the string). Controls are estimated based on probability functions of low level features, namely, the plucking instants (i.e. note onsets), the pitch and the distances of the fingers (both hands) to strings and frets. Note onsets are detected via audio analysis, the pitch is extracted from the score and distances are computed from 3D Euclidean Geometry. Results show that by combination of multimodal information, it is possible to estimate such a comprehensive set of control features, with special high performance for the fingering and plucked string estimation. Regarding the plucking finger and the pluck characteristics, their accuracy gets lower but improvements are foreseen including a hand model and the use of high-speed cameras for calibration and evaluation.

Keywords: Guitar, instrumental control, motion capture, audio analysis

1 Introduction

The acquisition of musical gestures and particularly of instrumental controls from a musical performance is a field of increasing interest with applications in acoustics (Schoonderwaldt et al., 2008), music pedagogy (Visentin et al., 2008), automatic music generation (Erkut et al., 2000; Maestre et al., 2010; Pérez-Carrillo et al., 2012), augmented performances (Bevilacqua et al., 2011;

Wanderley and Depalle, 2004) or performance transcription (Zhang and Wang, 2009) among others. On the classical guitar, a performer can produce very distinct sounds and adjust the timbre by the way the strings are plucked and pressed (Schneider, 1985). A comprehensive study of guitar controls is described by Scherrer (Scherrer, 2013). In that study, the principles of guitar playing in terms of controls are classified into left-hand (fingering) and right-hand (plucking). Fingering determines the pitch, as well as a set of more complex gestures such as vibrato, slurs, glissandi or damping. The main parameters during the plucking process are (a) the plucked string (b) the excitation with nail or flesh, (c) the plucking position, (d) the plucking velocity, (e) the displacement of the string during the pressure stage and (f) the angles of the pluck during the preparation and release stages.

Controlled measurement of every parameter is very difficult especially in a performance context. Reported methods in the literature generally focus on the estimation of a single parameter or a reduced set of them and methods are many times very intrusive. In this work we are able to extract a comprehensive set of control parameters from real performances by means of a multimodal combination of motion capture, sound analysis and information from a musical score. The extracted instrumental controls are the plucking instants (i.e. note onsets), the fingering (which bones of the left hand fingers are pressing which strings and frets), the plucked string, the plucking finger, the plucking position on the string and the plucking velocity and angles at the release stage.

Motion capture, based on high speed video cameras that detect the position of reflective markers, is used for tracking the position of the fingers and guitar strings. One problem inherent with optical motion capture is that of occlusion. Most motion capture involves the entire human body, which implies large body movements and the occurrence of occlusion is far less than that of small hand movements. Furthermore, it is the right hand of the guitarist that is extremely difficult to capture. In this work, motion capture is reinforced with audio analysis and the indications in the musical score.

Audio is recorded by means of a transducer measuring vibration on the guitar body and analyzed in order to detect the note onsets (Reboursière et al., 2012). There exist many different onset detection algorithms and they perform particularly well with the guitar due to the impulsive characteristics of string plucking (a comparison of different methods for guitar onset detection is reported in Reboursière et al. (2012)). Additionally, the fact that audio is measured as vibration of the guitar plate, makes onset detection even more accurate compared to a signal captured with a microphone, as the measured signal is not affected by room acoustics or sound radiation. Conversely to onset detection, pitch estimation algorithms such as de Cheveigné and Kawahara (2002) are not adapted to the guitar as note sustains are absent and note releases are very large making notes overlap in time. In order to have a robust estimation of the pitch, we use the note pitch information from the musical score as a ground truth, after an alignment to the audio signal.

The procedure for the parameter estimation proposed in this work is shown in Figure 1. The algorithm starts with (1) the estimation of the plucking instants and the pitch by onset analysis of the audio signal followed by (2) an alignment to the score. At each plucking instant (3) the distances from the fingers to the strings are computed and (4) the possible combination of fret and string is estimated from the pitch. By means of a probability function based on distances, the (5) most likely plucked string and (6) and the most likely plucking finger are estimated. Finally, from the selected string and finger the rest of the parameters (i.e. the plucking position, velocity and angles) are computed based on Euclidean geometry.

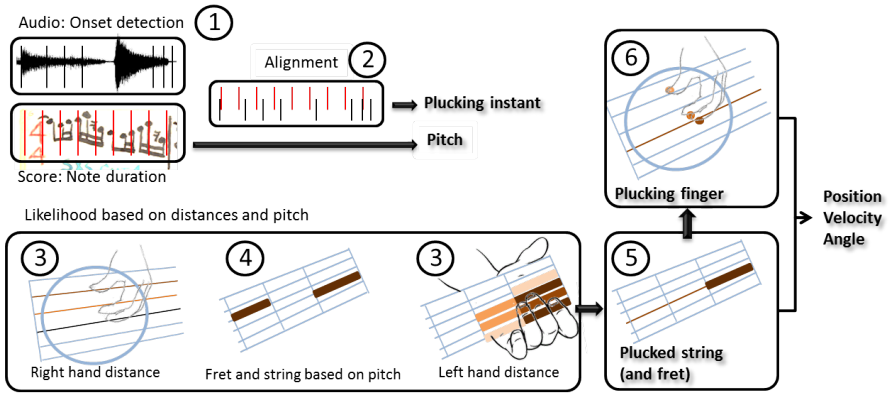

Fig. 1. Procedure for parameter estimation. It starts with (1) the estimation of the plucking instants by onset analysis of the audio signal and (2) pitch extraction by alignment of the score note onsets to the audio onsets. At each estimated plucking instant, (3) distances from fingers to the strings are computed. Given the pitch and the distances (4) a likelihood function is built in order to obtain (5) the most likely plucked string and fret and (6) the plucking finger. From the selected string and finger positions, the plucking position, velocity and angle are computed based on Euclidean geometry.

2 Literature Review

The acquisition of control parameters from musical performances is generally carried out either directly by measuring with sensors or indirectly by off-line analysis of an audio signal (Pérez-Carrillo and Wanderley, 2015; Wanderley and Depalle, 2004). In the case of the indirect acquisition, different methods allow for the extraction of the plucking position and fingering based on frequency-domain analysis techniques (Traube and Depalle, 2003; Traube and Smith, 2000) and time-domain approaches (Penttinen and Välimäki, 2004). Reboursière et al.

(2012) are able to detect left and right hand techniques (i.e. discriminating between left and right hand attacks and detection of right-hand palm-damping and harmonics) from audio analysis. Abesser and Lukashevich (2012) propose an algorithm that detects plucking parameters and expression styles for bass guitar playing. Scherrer and Depalle (2012) are able to estimate more complex features such as the plucking angle of release (AOR) by sound analysis informed with physical properties of the guitar.

The use of sensors allows to extract more parameters with higher accuracy and with the potential of acquisition in real time. The main reported techniques that provide a 3D representation of a live performance are based on mechanical, inertial, electro-magnetic or optical systems. Mechanical systems imply wearing a mechanical exoskeleton (Collins et al., 2010) that is very intrusive during performance. Inertial systems, most of them based on gyroscopes that measure rotational rates, have the disadvantage of being intrusive and of providing relative movement and not absolute position. Such systems have been used to track the movements of violin players (Linden et al., 2009). Electro-magnetic field (*EMF*) technology used for instance to measure violin bowing controls (Maestre et al., 2007; Pérez-Carrillo, 2009) is very accurate but generally intrusive and may have interferences with metallic objects and external magnetic fields. Finally, optical systems are widely used as they are generally low-intrusive and allow for highly accurate measurements. Burns and Wanderley (2006) studied how to capture the left-hand fingerings of a guitarist in real-time using low-cost video cameras. Their prototype system, using fret and string detection to track fingertips, was able to successfully identify chords and a series of notes. This was accomplished without the use of markers on the hands. Two acknowledged drawbacks of the system are that it can only capture finger movement on the first five frets of the guitar due to the choice of camera, and there is no way to address finger occlusion. Although preliminary, it was a first step to a possible real-time video automatic transcriber. Norton (2008) uses motion capture based on cameras detecting reflective markers, similar to the system employed for this research, to measure classical guitar performances. Heijink and Meulenbroek (2002) researched the complexity of left-hand classical guitar movements using an active optical motion capture system with four infrared light emitting diodes placed on the fingernails of the left hand, one on the left wrist, one each on the right index and middle fingers and three on the body of the guitar. Chadefaux et al. (2012) used high speed video cameras to manually extract features during plucking of harp strings.

Other types of measuring techniques can also be found, including methods based on capacitive sensing (Guaus and Arcos, 2010) to capture left-hand fingering and indirect acquisition from audio (Penttinen and Välimäki, 2004; Scherrer and Depalle, 2012; Traube and Depalle, 2003; Traube and Smith, 2000). The selection of a measuring system is largely determined by the objectives of the research. In this work the main objective is to measure hand controls from real performances with high accuracy and no (or very low) intrusiveness, which de-

termined the choice for a high speed camera system that captures the position of small and ultra-light reflective markers as in Norton (2008).

3 Multimodal Data Acquisition

A multimodal database of guitar performances was recorded in order to test and evaluate the algorithms. The database is composed of ten musical fragments with an average duration of around one minute, performed by two different guitarists. The database contains the audio, 3D motion data, information from the musical score (note onset, note offset, pitch and the ground truth for the parameters plucked string, plucking finger, fret and left-hand fingering). Audio and Motion streams were recorded in different computers synchronized by means of a world clock generator that controls the sampling instants of the capturing devices and sends a SMPTE signal of 25Hz that is saved as timestamps with the data. Audio to motion alignment consists of simply aligning the SMPTE timestamps.

3.1 Audio recording

Audio is recorded with a contact microphone that captures the vibration of the guitar body. The captured signal is better adapted for audio analysis than that of a microphone as it is not affected by room acoustics or sound radiation. The audio stream is segmented into notes by means of onset detection based on the complex domain algorithm (Duxbury et al., 2003). Conversely to pitch detection, onset detection in guitar playing is very accurate as notes are plucked, which implies a high energy peak at note onsets (Reboursière et al., 2012).

3.2 Musical score

The musical score provides the nominal pitch, onsets and offsets of the notes, which are aligned with the ones detected in the audio. Figure 2 shows an example using wavesurfer software [4] of an audio segmentation showing three *label* tracks that indicate the score onsets and pitch (track *.notes*) along with the ground truth for the string (track *.string*) and the fret (track *.fret*) to be played.

3.3 Motion Capture

Motion capture is used to track the position of finger bones and articulations as well as the guitar strings. The capture is optical by means of *Qualysis* [5] high speed video cameras that detect the position of reflective markers. The main problem with such optical MoCap systems is marker occlusion. Each marker needs to be identified by at least three cameras placed at different angles and planes in order to correctly determine its 3D coordinates. In order to achieve a

[4] http://sourceforge.net/projects/wavesurfer/
[5] http://www.qualisys.com/

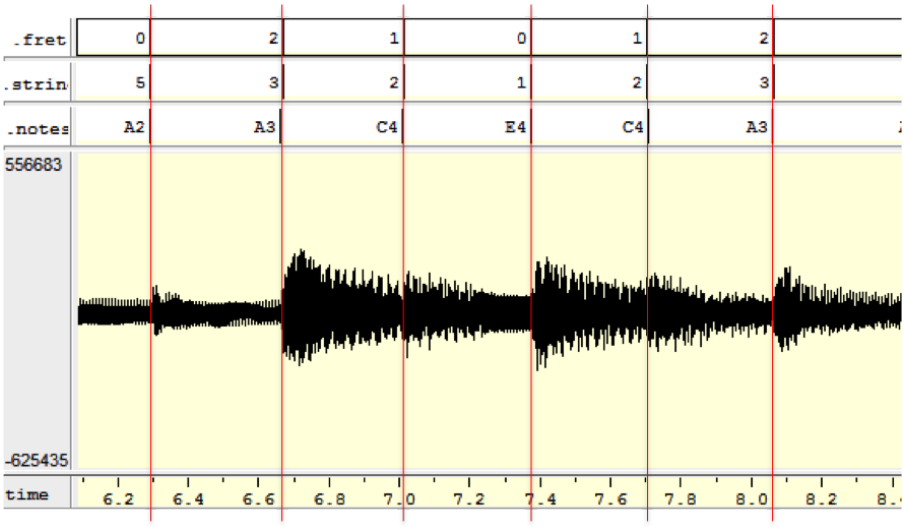

Fig. 2. Visualization with *wavesurfer* software of a fragment of a performed score showing the waveform segmented into notes together with three *label* tracks with the ground truth from the musical score information (the fret, the string and the note pitch). The segmentation is obtained after the onset detection and alignment of the score to the detected onsets.

correct identification of the markers, it is necessary a careful placement of the cameras, the use of models for the hands and the guitar, and if necessary, the manual cleaning of the data, i.e. assigning the appropriate labels to incorrectly identified and non-identified markers.

3.4 Hand tracking.

The motion of the fingers is followed by attaching a marker at each finger articulation as shown in Figure 3. Left-hand fingering estimation is very accurate as marker occlusion is low. Conversely, the right hand of the guitarist is extremely difficult to capture as it implies small hand movements and it gets especially complicated on the markers attached to the fingernails due to the particular way of playing the guitar (when plucking, the nails face the guitar body). The *Qualisys* software includes algorithms for the definition of skeleton models that are trained with recordings of the hands moving smoothly and allow for the automatic identification and reconstruction of lost markers from skeleton objects. Proceeding this way, we achieve a higher rate of correctly recognized marker trajectories.

3.5 String Coordinates.

The position of the strings is determined by the definition of a guitar Rigid Body (*RB*). A *RB* is a six degrees-of-freedom (6DOF) rigid structure defined by the position of a set of markers and associated with a local system of coordinates (SoC) with a corresponding 3D position and orientation with respect to the global SoC. The position of the markers is constant relative to the local SoC and their global coordinates can be obtained by a simple rotation and translation from the local to the global SoC. The guitar *RB* is built by placing markers at each string-end as well as *reference* markers attached to the guitar body. The markers at the string-ends can not remain attached during a performance as it would be very intrusive, so they are defined as virtual markers. Virtual markers are only used for calibration to define the SoC structure. During the actual tracking, virtual markers are reconstructed from the reference markers.

Special care has to be taken when attaching the reference markers to the guitar body as the plates are reflective to light, causing interferences with the markers. In order to avoid these unwanted reflections, the markers are placed on the edge of the guitar body and outside the guitar plates by means of antennas (prolongations attached to the body). Five reference markers were used and they were placed as shown in Figure 3 (blue markers). The tracking of the position of the strings following this procedure achieves a nearly 100% of correctly reconstructed frames.

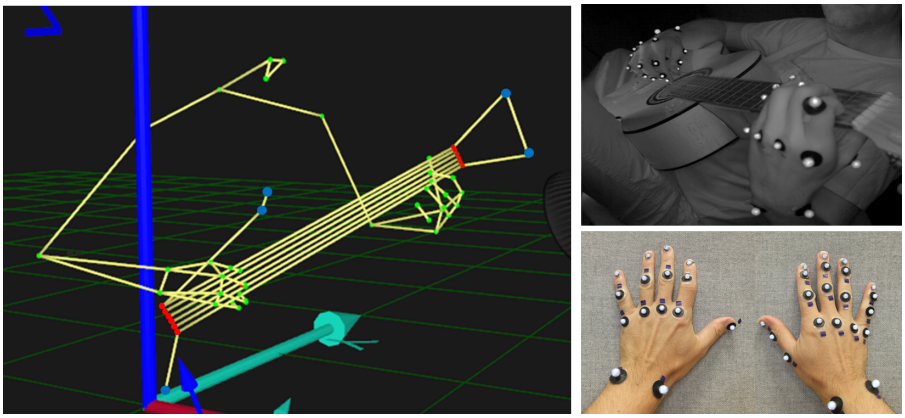

Fig. 3. To the left, a 3D visualization of the motion capture. Blue dots are the guitar auxiliary markers, which are used for the capture of the string positions, red dots represent the virtual markers on the strings ends, and green markers are the joints in the body of the performer. At the top-right, one of the capture frames with a camera. At the bottom-right, the position of the markers in the hands.

4 Low level parameter Computation

The estimation of plucking and fingering features is based on the extraction of low level parameters, namely, the plucking instant t_0, the fundamental frequency (or pitch) f_0 and the finger distances (both hands) to the strings.

4.1 Plucking instant t_0 and fundamental frequency f_0.

The plucking instants are determined from the audio by means of a note onset detection algorithm (Duxbury et al., 2003) followed by an alignment and match to the note start times in the musical score.

Detected onsets (o^A) are aligned and matched with note starts in the score (o^S) in order to have a robust estimation that discards false detected onsets and allows to restore non detected ones. Once the score is aligned with the audio, we can extract the note pitch f_0 directly from the score.

The alignment is performed in two steps. First, a window indicating the approximate start and the end of the performance in the recording must be found and then, we proceed to the alignment inside the window. The first step is necessary because audio recordings start and finish with a silence of undetermined duration and the window allows to discard false o^A due to noises outside the actual performance.

The boundaries of the performance are found by computing a smoothed audio energy envelope using a time-sliding window of length 100 frames ($24ms$ at a sample rate of $42100Hz$) and by defining a minimum hearing threshold (h_θ) as shown in Figure 4. The procedure results in an envelope with high values of energy around the real performance. The threshold is used to determine the start and end of the window. The beginning of the performance will be at the first o^A that lays inside the window and the last note onset is at the last o^A inside the window. Once the boundaries of the performance are determined, the total score duration is stretched to fit the actual performance duration.

The second step consists of matching o^S (from the stretched score) to o^A. For this we define an algorithm that for each o_i^S looks for its closest o_j^A. If the found o_j^A is closest to o_i^S than to o_{i+1}^S, both o_i^S and o_j^A are matched together and the new time is updated in the stretched score (o_i^S is set equal to o_j^A and distances in the score are delayed or advanced o_i^S-o_j^A). If no match is found for o_i^S we consider that we have detected a missing onset:

```
for i=2:length(oS)  {
    j=findClosest(oS(i), oA)
    if(i==finClosest (oA(j), [oS(i),oS(i+1)]))
        match(i,j)
}

function match(i,j) {
    diff=oS(i)-oA(j)
    oS(i).time=oA(j).time
    for k=i:length(oS(i))
    oS(k).time=oS(k).time+diff
}
```

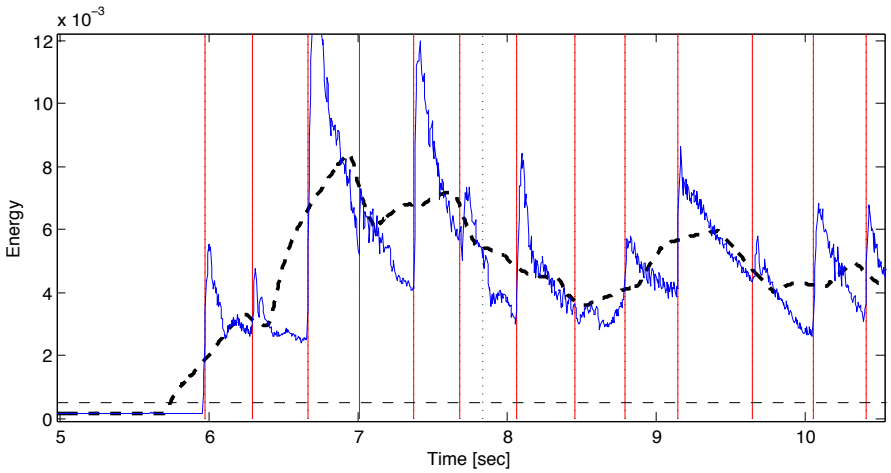

Fig. 4. Score to Audio alignment. Energy in the audio is represented as continuous line in blue and the smoothed energy envelope is depicted with a thick dashed line. The hearing threshold is the straight dashed line close to zero and the aligned onsets are the vertical red lines. The boundaries of the recording are found inside the smoothed energy envelope and over the hearing threshold.

4.2 Finger distance to the strings

Distance from the markers in the fingers to the strings is computed as the average distance in a small window around the note onsets. A distance of zero would indicate that a finger is in contact with the string, but in practice, there is an offset in the distance as markers are placed on the back of the hand and this offset has not been calibrated. However, the method does not need to know the exact distances as it is based on the probability of being in contact with a string.

Left Hand. The left-hand presses the strings at specific frets, determining this way the unique pitch that can be played at each string. A string can be pressed not only with the finger tips but with any part of the finger and the pressing position (and therefore the pitch) is discretized by the position of the frets. For this reason, both fingers and strings are represented as segments. A finger-segment is determined by two consecutive markers in the same hand and a fret-segment is defined as the string segment between two consecutive frets. The coordinates of the finger-segments are obtained directly from the measured left-hand marker positions and the position of the string-segments are to be found along the string lines at distances from the nut x_N and the bridge x_B given by,

$$\begin{aligned} x_N(j) &= \frac{x_B(j-1)}{k} + x_N(j-1), \\ x_B(j-1) &= x_S - x_N(j-1), \end{aligned} \tag{1}$$

where x_S is the string length, $k = 17.817$ is a known constant that relates distances among frets and $j = 1..24$ is the fret number.

At each plucking instant (determined by the note onsets) we compute the distances from fret-segments to any of the finger-segments and we define the left hand distances function $d_L(s, f, b)$, where s is the string, f the fret-segment and b the finger-segment or bone. Distances are computed as the shortest Euclidean distance between two line segments (Eberly, 2000). Additionally we define,

$$\begin{aligned} d_L(s, f) &= \min_b d_L(s, f, b), \\ d_L(s) &= \min_f d_L(s, f). \end{aligned} \tag{2}$$

Right hand. In a similar way to the left hand, we define the right hand distances function $d_R(s, fg)$, being fg the finger, as the distances from the finger-nails to the strings, so distances are computed as the shortest distance between a point (finger nail) and a line (string) (Eberly, 2000). Additionally we define,

$$d_R(s) = \min_f d_R(s, fg). \tag{3}$$

4.3 Likelihood functions

Control parameters are estimated based on likelihood functions obtained from the previously described low level features. Let $L_P(s, f|pitch)$ be the likelihood of pressing a fret f and plucking a string s given the *pitch*. As the confidence of the pitch is very high (it is the ground truth determined from the score) we can use a binary function to discard impossible combinations of $frets$ and $strings$. Its value is set to one at the combinations where the *pitch* can be produced and zero otherwise. This function restricts the possible *frets* to at most one per *string*, so we can also define $L_P(s|pitch) = L_P(s, f|pitch)$.

To determine the likelihood that a finger is touching a string, we define a function θ that maps the distances d_L and d_R ($\in \mathbb{R}$) to likelihood ($\in \mathbb{R} \in [0..1]$). The range of distances of interest (where the finger could be considered to be in contact with the string or very close) is from a few millimeters below the string (negative distance) to around 1cm above. We therefore need an asymptotic function that tends to 0 for large values of the distance and tends to 1 for small (positive and negative) distances. For distances around 0.5 cm, for which it is not clear if the finger is in contact with the string, the likelihood should be around 0.5. The function is defined as $\theta(d) = \arctan(d * 10)/\pi + 1/2$, where d is the distance, and therefore,

$$\begin{aligned} L_L &= \theta(d_L), \\ L_R &= \theta(d_R). \end{aligned} \tag{4}$$

In addition, we need to define the likelihood of playing an open string (i.e. not stopping the string with the left hand fingers, $f = 0$). It is computed as $L_L(s, 0) = 1 - \sum_f L_L(s, f)$, that is, one minus the sum of likelihoods of being pressing the other frets in the same string.

5 Control Parameter Estimation

From the low level features we estimate in a first step the plucked string, the fret, the left-hand segment pressing that string at that fret and the plucking finger. In a second step, the plucking position, the plucking velocity and the plucking angles are computed using the coordinates of the string and plucking finger.

5.1 Plucked String (*string*)

It is the most likely string to be plucked at a note onset. The likelihood that a string is being played (L_S) is determined as a function of the pitch and the distances to the strings of the left and right hand fingers, d_L and d_R.

$$\begin{aligned} string &= \max_s(L_S(s)) \\ L_S(s) &= L_P(s) \times L_L(s) \times L_R(s), \end{aligned} \tag{5}$$

5.2 Plucking finger (*finger*)

Once we know the plucked *string* (string with maximum L_S), it is straightforward to obtain the right-hand *finger* that plucks the *string*. It is the closest finger to the *string*, that is, the finger that maximizes the likelihood function L_R given the string:

$$finger = \max_{s=s_i} \theta(d_R(s, fg)). \tag{6}$$

5.3 Fingering (*fret* and *bone*)

Ir refers to the position of left-hand fingers on the strings. Once the *string*, and the left hand finger distances have been estimated we can define the *fret* and *bone* as

$$\begin{aligned} fret &= \max_{s=s_i} \theta(d_L(s, f)) \\ bone &= \max_{s=s_i, j=j_j} \theta(d_L(s, f, b)). \end{aligned} \tag{7}$$

5.4 Plucking Position (C)

The plucking position, velocity and angle are computed from the *string* and *finger* based on Euclidean Geometry as shown in Figure 5. Be A and B the plucking *string* ends, and P the position of the plucking *finger*, the plucking position is the point of contact of the *string* and *finger*. Due to the position of the markers on the back of the hand, this may not be zero, so we define the point C as the point in the string that is closest to P. This point computed as the projection of the point P on the string,

$$C = AB \cdot AP * \widehat{AB}. \tag{8}$$

Given the parametric definition of a line in 3D, we can define the point C as

$$C = A + tAB, 0 < t < 1, t \in \mathbb{R}. \tag{9}$$

From eqs. 8 and 9 we derive the value of $t = (AB \cdot AP)/|AB|^2$ for the point C. The plucking distance to the bridge (i.e. plucking position) is just the length of the segment $|AC|$, which corresponds to the value of the dot product $AB \cdot AP$.

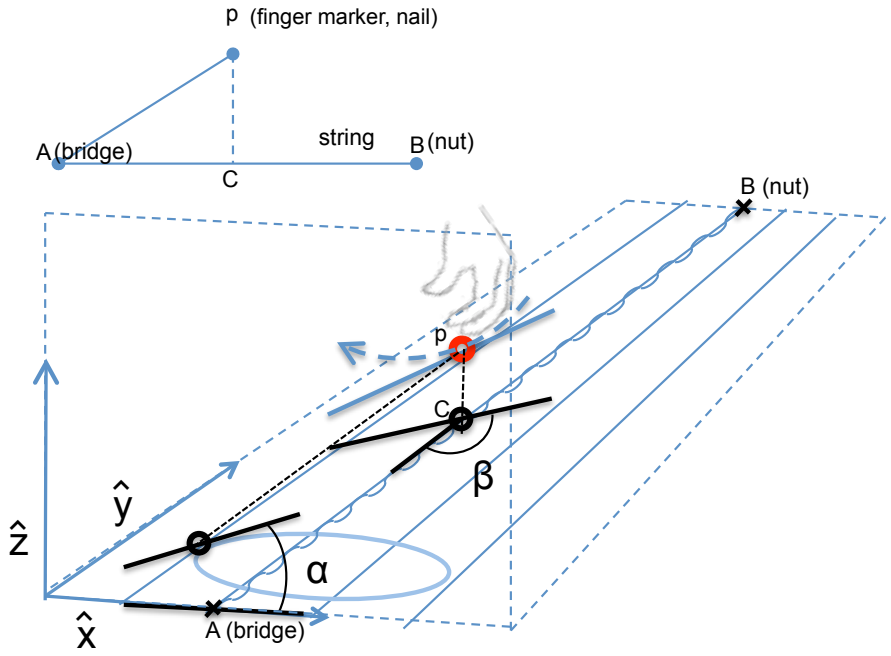

Fig. 5. Computation of plucking position, velocity and angles based on Euclidean Geometry.

5.5 Plucking Velocity (v)

The velocity is computed as the derivative of the marker trajectory at the plucking instant.

5.6 Plucking Angles (α and β)

Two angles are estimated as shown in Figure 5. The angle α is the angle between the projection of the tangent of the finger trajectory at the plucking instant on the plane xz and the axis x. The second angle β is the angle between the projection of the tangent on the plane xy and the axis y. In order to simplify geometric computations, all points (string-end positions and finger trajectory around the plucking instant) are expressed relative to the local axis of coordinates (xyz), where the vector $\widehat{x}$ is the unitary vector in the direction from the beginning of the sixth string to the beginning of the first string, vector $\widehat{y}$ is the unitary vector in the direction of the playing string towards the frets, and $\widehat{z}$ is obtained as the cross product $\widehat{z} = \widehat{x} \times \widehat{y}$.

Any point P in global coordinates can be expressed in local coordinates P' by rotation (R) and traslation (T):

$$P' = (P - T) * R^{-1}, \tag{10}$$

where the traslation is given by the coordinates of the vector that connects the origin of coordinates from the global to the local system and the rotation is expressed as an Euler rotation matrix [6], which is computed from the rotation angles of the local coordinate axis with respect to the global axis. By using local coordinates, the computation of the projection of the finger trajectory $F(t)$ on the planes xy and xz (i.e. $F_{xy}(t)$ and $F_{xz}(t)$) becomes straight forward as they merely correspond to the coordinates of the axis. The tangents to the projected trajectories at the plucking instant have a slope equal to the derivative of the projected trajectories at that point, and this slope is also the tangent of the angles, so that:

$$\begin{aligned} \alpha &= \arctan \frac{dF_{xy}(t)}{dt}, t = t_0 \\ \beta &= \arctan \frac{dF_{xz}(t)}{dt}, t = t_0. \end{aligned} \tag{11}$$

6 Results

The performance of the presented method depends on a correct onset detection and marker identification. We can assume a very high rate of onset detection that is further improved by the alignment to the score (see section 4) and marker identification rates are specified in Table 1, expressed as percentage of correctly identified markers relative to all frames as well as to the plucking instants. Identification of string markers achieves a rate of nearly 100% but finger markers results are very variable depending on the marker.

[6] http://mathworld.wolfram.com/EulerAngles.html

Table 1. Percentage of correct marker identification taking all frames into account (*all frames*) and only frames at plucking instants (*at plucks*). The three joints in the left hand follow an order from closer to the palm towards the fingernail.

Marker	% *all frames*	% *at plucks*
String-ends	99	100
Left-index (3 joints)	99, 99, 83	100, 100, 87
Left-mid (3 joints)	99, 99, 60	100, 99, 58
Left-ring (3 joints)	100, 99, 77	100, 100, 71
Left-small (3 joints)	99, 99, 30	100, 100, 21
Right-thumb (nail)	99	100
Right-index (nail)	79	76
Right-mid (nail)	82	75
Right-ring (nail)	70	64
All 4 right hand markers	50	41

6.1 Left Hand

Left hand markers have a higher rate of correct identification as, in general, they are always visible to the cameras. Additionally, fingering (i.e. left hand) is highly reinforced with the pitch, which largely determines the *string* and the *fret*, so even if a marker is lost, the closest finger is 100% correctly estimated.

6.2 Right Hand

Regarding the right hand, only one marker per finger is considered, which is the one placed on the fingernail. It is especially these markers during the note onsets that are difficult to track due to occlusions. An average of 75% of the markers are correctly identified but only around 41% of the note onsets have all four markers (the four nails) and only 64% of the plucks with the ring-finger are detected.

Two improvements have been added to the tracking of the right hand. First, markers are searched within a small window around the plucking instants, which allows to increase the chances that the plucking instant frame has every marker. For instance, in Table 2 we can see the difference of correctly estimated *plucking fingers* using a window of 3 frames (46%) and a window of 21 frames (58%). Second, marker trajectories are interpolated in order to fill missing gaps. Different types of interpolation were compared. If the marker trajectory gaps are small the type of interpolation does not affect the estimation but if the gap is big, the selection of the interpolating algorithm becomes very important. The best performing interpolation is a *piecewise cubic hermite spline* achieving rates for string and fret estimation of 100% and around 75% for the right and left fingers. In Table 2 we can find a summary of the results with different windows and interpolation algorithms.

Table 2. Percentage of correctly estimated plucking *finger* using different window sizes w (which allows to look for missing markers in frames around a plucking instant) and different marker interpolation algorithms: no interpolation (*no interp.*), cubic splines (*spline*), cubic smoothing splines (*csaps*), nearest neighbor (*nearest*), linear and piecewise cubic hermite polynomial (*pchip*).

	% plucking finger
no interp., w=3	46
no Interp., w=21	58
spline, w=3	64
csaps, w=3	64
nearest, w=3	73
linear, w=3	75
pchip, w=3	78

6.3 Position, Velocity and Angles

The error of the characteristics of the pluck (position, velocity and angles) is estimated by using notes where the plucking fingernail markers are correctly identified. For these notes, plucking parameters are computed and their values compared to an estimation of the same pluck, but with missing marker position information. The trajectory of the marker is removed at a window of 10 frames around the note onset and a new trajectory is estimated through *pchip* interpolation. Such an evaluation results in a Correlation Coefficient of 0.83 for the plucking position, 0.75 for the velocity and 0.73 and 0.72 for the plucking angles α and β respectively.

7 Conclusion

In this study we showed a method for the extraction of a comprehensive set of control parameters from guitar performances based on hand-motion capture and supported by audio analysis and information on the score. The extraction of control parameters will allow for future research on the interaction between the guitar and the player, the relationship between gestures and sound and the analysis of different playing techniques to mention a few fields.

Results show that by combination of multimodal information, it is possible to estimate such a comprehensive set of control features, with special high performance for the fingering and plucked string estimation. The two most relevant contributions of the method are (1) the combination of multimodal data. Solely based on motion capture it would be very complicated to detect the correct onsets or plucking instants. Additionally, the availability of the ground truth pitch highly reinforces the detection of string, fingering and fret; (2) interpolation of marker trajectories and the use of a plucking window makes marker identification more robust and boosts the estimation rates.

Several improvements are foreseen for the future. (1) Develop an algorithm for pitch extraction from audio adapted to the guitar to avoid the use of a musical score; (2) build a 3D flexible object model of the hand and a hand motion grammar in order to restrict the possible positions of the flexible object and be able to reconstruct missing markers from the position and angles of the identified joints; (3) Use of a finger-tip model in order to calibrate the position of the real nail and flesh with respect to the marker; (4) The use of high speed video cameras to calibrate and evaluate data from the motion capture and deepen the analysis of the plucking process (start, contact and release) as done in Chadefaux et al. (2012).

Acknowledgments

A. Perez-Carrillo was supported by a Beatriu de Pinos grant 2010 BP-A 00209 by the Catalan Research Agency (AGAUR) and J. Ll. Arcos was supported by ICT -2011-8-318770 and 2009-SGR-1434 projects.

Bibliography

Abesser, G. S. J. and Lukashevich, H. (2012). Feature-based extraction of plucking and expression styles of the electric bass. In *Proceedings of the ICCASP Conference*, Kyoto, Japan.

Bevilacqua, F., Schnell, N., Rasamimanana, N., Zamborlin, B., and Guédy, F. (2011). Online gesture analysis and control of audio processing. In Solis, J. and Ng, K., editors, *Musical Robots and Interactive Multimodal Systems*, volume 74 of *Springer Tracts in Advanced Robotics*, pages 127–142. Springer Berlin Heidelberg.

Burns, A.-M. and Wanderley, M. M. (2006). Visual methods for the retrieval of guitarist fingering. In *Proceedings of NIME*, pages 196–199, IRCAM, Paris, France.

Chadefaux, D., Carrou, J.-L. L., Fabre, B., and Daudet, L. (2012). Experimentally based description of harp plucking. *The Journal of the Acoustical Society of America*, 131(1):844–855.

Collins, N., Kiefer, C., Patoli, Z., and White, M. (2010). Musical exoskeletons: Experiments with a motion capture suit. In *New Interfaces for Musical Expression*, Sydney, Australia.

de Cheveigné, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.

Duxbury, C., Bello, J. P., Davies, M., and Sandler, M. (2003). Complex domain onset detection for musical signals. In *In Proc. DAFx*, Queen Mary University, London, UK.

Eberly, D. H. (2000). *3D game engine design: a practical approach to real-time computer graphics*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Erkut, C., Välimäki, V., Karjalainen, M., and Laurson, M. (2000). Extraction of physical and expressive parameters for model-based sound synthesis of the classical guitar. In *Audio Engineering Society Convention 108*.

Guaus, E. and Arcos, J. L. (2010). Analyzing left hand fingering in guitar playing. In *Proc. of SMC*, Universitat Pompeu Fabra, Barcelona, Spain.

Heijink, H. and Meulenbroek, R. (2002). On the complexity of classical guitar playing: Functional adaptations to task constraints. *Journal of Motor Behavior*, 34(4):339–351.

Linden, J. v. d., Schoonderwaldt, E., and Bird, J. (2009). Towards a real-time system for teaching novices correct violin bowing technique. In *IEEE International Workshop on Haptic Audio visual Environments and Games*, pages 81–86.

Maestre, E., Blaauw, M., Bonada, J., Guaus, E., and Pérez, A. (2010). Statistical modeling of bowing control applied to sound synthesis. *IEEE Transactions on Audio, Speech and Language Processing. Special Issue on Virtual Analog Audio Effects and Musical Instruments.*

Maestre, E., Bonada, J., Blaauw, M., Pérez, A., and Guaus, E. (2007). Acquisition of violin instrumental gestures using a commercial EMF device. Copenhagen, Denmark.

Norton, J. (2008). *Motion Capture to Build a Foundation for a Computer-Controlled Instrument by Study of Classical Guitar Performance*. PhD thesis, Department of Music, Stanford University.

Penttinen, H. and Välimäki, V. (2004). A time-domain approach to estimating the plucking point of guitar tones obtained with an under-saddle pickup. *Applied Acoustics*, 65(12):1207 – 1220.

Pérez-Carrillo, A. (2009). *Enhancing Spectral Synthesis Techniques with Performance Gestures using the Violin as a Case Study*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain.

Pérez-Carrillo, A., Bonada, J., Maestre, E., Guaus, E., and Blaauw, M. (2012). Performance control driven violin timbre model based on neural networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):1007 –1021.

Pérez-Carrillo, A. and Wanderley, M. (2015). Indirect acquisition of violin instrumental controls from audio signal with hidden markov models. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 23(5):932–940.

Reboursière, L., Lähdeoja, O., Drugman, T., Dupont, S., Picard-Limpens, C., and Riche, N. (2012). Left and right-hand guitar playing techniques detection. In *NIME*, University of Michigan, Ann Arbor.

Scherrer, B. (2013). *Physically-informed indirect acquisition of instrumental gestures on the classical guitar: Extracting the angle of release*. PhD thesis, McGill University, Montréal, QC.

Scherrer, B. and Depalle, P. (2012). Extracting the angle of release from guitar tones: preliminary results. In *Proceedings of Acoustics*, Nantes, France.

Schneider, J. (1985). *The contemporary guitar*, volume 5 of *New Instrumentation*. University of California Press, Berkeley.

Schoonderwaldt, E., Guettler, K., and Askenfelt, A. (2008). An empirical investigation of bow-force limits in the Schelleng diagram. *AAA*, 94(4):604–622.

Traube, C. and Depalle, P. (2003). Deriving the plucking point location along a guitar string from a least-square estimation of a comb filter delay. In *IEEE Canadian Conference on Electrical and Computer Engineering.*, volume 3, pages 2001 – 2004 vol.3.

Traube, C. and Smith, J. O. (2000). Estimating the plucking point on a guitar string. In *Proceedings of COST-G6 conference on Digital Audio Effects*, Verona, Italy.

Visentin, P., Shan, G., and Wasiak, E. B. (2008). Informing music teaching and learning using movement analysis technology. *International Journal of Music Education*, 26(1):73–87.

Wanderley, M. M. and Depalle, P. (2004). Gestural control of sound synthesis. In *Proc. of the IEEE*, pages 632–644.

Zhang, B. and Wang, Y. (2009). Automatic music transcription using audio-visual fusion for violin practice in home environment. Technical Report TRA7/09, Shool of Computing, National University of Singapore.