Visual Analysis of Fingering for Pedagogical Violin Transcription

Bingjun Zhang Jia Zhu Ye Wang Wee Kheng Leow
Department of Computer Science, National University of Singapore
E-mail:{bingjun,zhujia,wangye,leowwk}@comp.nus.edu.sg

ABSTRACT

Automatic music transcription, in spite of decades of research, remains a challenging research problem. The traditional audio-only approach has yet to achieve a satisfactory performance for any computer-aided pedagogical system. Inspired by the high correlation between violin playing techniques (fingering, bowing) and the played acoustic notes, this paper presents a first attempt in visual analysis of violin fingering to compensate for the difficulties in audio-only music transcription. This is achieved by a robust multiple finger tracking algorithm and a string detection method that extract press, release, and fingertip position from the fingering video and automatically translate the fingering information into the played acoustic note, i.e., onset, offset, and pitches. Experimental results reveal high correctness in multiple finger tracking and string detection, thus paving the way for an improved audio-visual violin transcription system.

Categories and Subject Descriptors

H.5.5 [**Sound and Music Computing**]: Methodologies and techniques, Systems; I.2.10 [**Vision and Scene Understanding**]: Motion, Shape, Video analysis; I.4.8 [**Scene Analysis**]: Motion, Tracking

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Music transcription, multiple finger tracking, automatic note inference, violin fingering chart

1. INTRODUCTION

In recent years, computer-aided pedagogical systems have been developing rapidly to replace or significantly reduce human efforts in various education scenarios [1, 2]. For music education, an audio-only music transcription system has difficulties in accurately detecting onset, offset and pitches for every note, which is the basic task of music transcription. In particular, for pitched non-percussive (PNP) sounds such as from the violin, the onset and the offset of a note often change gradually over a prolonged period of time. Therefore, it is difficult to pinpoint the exact time instances at which the onset and the offset occur [3] (Figure 3). Moreover, when two notes of the same pitch are played consecutively, the change from the offset of the first note to the onset of the second note is very subtle and so, it is very difficult to detect the onset/offset using audio signal only. Multi-pitch estimation also remains a research challenge despite some recent progress [4].

Inspired by the high correlation between violin playing techniques (fingering, bowing) and the played acoustic notes, visual information of fingering and bowing can be used to complement audio-only methods to enhance transcription performance. We tackle the difficulty for onset/offset detection from the gradual change of audio signal by extracting finger press and release events from fingering video, and solve the problem of onset/offset detection of consecutive notes with the same pitch by tracking bowing. Multi-pitch estimation can also be enhanced by extracting fingertip positions from the fingering video. Besides improving transcription performance, the audio-visual approach can also provide useful visual information, such as fingering and bowing trajectories and playing gestures, to the player as learning feedback. Due to space limitation, in this paper, we focus on the automatic visual analysis of violin fingering video for improved audio-visual violin transcription. To our knowledge, this is also the first attempt to use bare finger tracking for music transcription.

This paper presents a method for inferring the elements of the played acoustic notes (onset, offset, and pitches) by robust multiple finger tracking, string detection and physics-based automatic note inference (Section 3).

Experimental results (Section 4) reveal high correctness in our method for multiple finger tracking and string detection, thus paving the way for an improved audio-visual violin transcription system.

2. RELATED WORK

Several methods have been proposed to solve the automatic finger tracking problem, such as [5, 6, 7]. In [5], fingertips are tracked in bare hand video with flat palm, which is not applicable to the violin fingering analysis as the fingers are always bent while playing a violin. An articulated hand tracking method is proposed in [6] with the condition

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

that the palm and the fingers are not occluded. However, this is clearly not the case for violin fingering since the palm is always occluded by the neck of the violin.

The work that is most similar to ours is [7] on guitarist finger tracking. In [7], Burns and Wanderley used Hough circle transform to detect fingertips based on the clear fingertip edges in the video. Our experiments show that this method will fail if the finger moves away from the finger board and fingertip edges cannot be easily detected, which happens frequently in violin video. Furthermore, this method is not robust against the noise near the fingertips.

In summary, none of the previous work is suitable for violin finger tracking either because of their constraints to specific application scenarios or their poor performance. In this paper, we propose a robust approach to multiple finger tracking in violin fingering video based on Condensation algorithm [8] by employing joint finger model dynamics and Gaussian skin color model.

3. VISUAL ANALYSIS OF FINGERING

Visual analysis of fingering is achieved by tracking the four fingers of the violinist's left hand and detecting the four strings (string E, A, D and G) of the violin to obtain the fingering events from each frame of the fingering video. A bird's eye view of the violin finger board (Figure 4) is selected to capture the necessary information for visual analysis of violin fingering. The visual analysis consists of five stages which are discussed in the following sections.

3.1 Motion Compensation

Global motion compensation technique is applied to reduce the global translation of the fingers and the violin. This decreases the complexity of multiple finger tracking and string detection. By referring to the first frame, motion vectors of subsequent frames are computed by finding the best match between each frame and the first frame. Then, each frame is translated in the opposite direction of the motion vector to remove global translation.

3.2 Multiple Finger Tracking

The four fingers of the violinist's left hand are tracked simultaneously using Condensation algorithm [8] in which joint finger model dynamics and Gaussian skin color model are employed.

3.2.1 Joint Finger Model Dynamics

Each finger is modeled by a closed B-spline curve with eight control points, q_1 to q_8 (Figure 1), which form a feature vector $Q = (x_{q_1}, ..., x_{q_8}, y_{q_1}, ..., y_{q_8})^T$ that captures the articulated finger contour. A finger shape space is constructed separately for each finger as follows. N training samples, $Q_1, ..., Q_N$, consecutive in time are manually collected. The difference vector between Q_i and Q_1, for each i, is computed. Principal Components Analysis (PCA) is applied to the difference vectors to obtain the shape space W that transforms the difference vector to a shape vector, F_k [9]:

$$Q_k - Q_1 = W F_k \ . \tag{1}$$

The joint model dynamics of the four fingers are formulated as a second-order auto-regressive process and learned from the shape vectors F_{jk}, $j = 1, ..., 4$, and $k = 2, ..., N$,

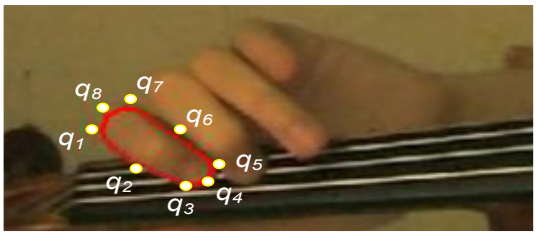

Figure 1: The finger model. The finger contour is modeled by a closed B-spline curve with eight control points.

where j is the finger number:

$$\begin{aligned} &X_t = A_1 X_{t-1} + A_2 X_{t-2} + B_0 w_t \\ &P(X_t | X_{t-1}, X_{t-2}) \propto \\ &\quad \exp\left\{-\tfrac{1}{2}\|B^{-1}(X_t - A_1 X_{t-1} - A_2 X_{t-2})\|^2\right\} \end{aligned} \tag{2}$$

where $X_t = \left[\ F_{1t}^T\ F_{2t}^T\ F_{3t}^T\ F_{4t}^T\ \right]^T$. The joint finger model dynamics include inter-finger constraints, such as finger order, no overlap between two fingers, etc., which are important for correct tracking of the finger contours. As long as there are enough training samples, Eq. 2 will be able to accurately model the articulated finger contour dynamics during violin playing.

3.2.2 Finger Measurement Density

The finger measurement density $P(Z|X)$ is estimated as follows. First, Canny edge detector is applied to extract edges from the input image. Edge pixels with colors similar to a pre-computed Gaussian skin color model are identified. The directions of the edges are also computed.

Once the set Z of edges is computed, given a predicted shape vector X, the measurement density is computed by searching along S normals of the predicted contour of each finger [9] as:

$$P(Z|X) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{S}\min(\|p_i - z_i\|, \rho)\right\} \tag{3}$$

where p_i and z_i are respectively the pixel position on the predicted contour and the signed edge along the i-th normal, ρ is a constant penalty, and σ is the standard deviation.

3.2.3 Iterative Estimation

Based on the above joint finger model dynamics and measurement density, Condensation algorithm [8] is applied to iteratively estimate the current finger contours:

$$P(X_t|Z_t) \propto \prod_{j=1}^{4} P(Z_{jt}|X_{jt}) P(X_t|X_{t-1}, X_{t-2}) \tag{4}$$

where measurement densities of the four fingers are assumed to be mutually independent.

The contour of each finger in every frame is tracked. For each frame, the fingertip position, $T_j = (x_j^t, y_j^t)$, of each finger contour is computed as the average position of the control points q_3, q_4, q_5 of the B-spline curve of the contour.

3.3 String Detection

As there is good contrast between the white strings and the black finger board, and the strings are straight except

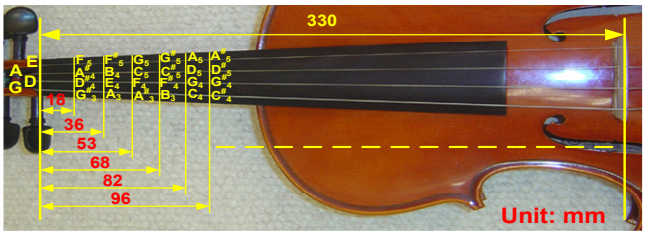

Figure 2: Violin fingering chart built on the physics of the violin and the vibrating string.

at the nut and the bridge of the violin, it is relatively easy to detect the starting point $P_j = (x_j^p, y_j^p)$ (at the nut) and ending point $P_j' = (x_j^{p'}, y_j^{p'})$ (at the bridge) of each string by first applying Hough line transform [10] and then searching for the turning points along the detected line track. Figure 4 illustrates some sample results of string detection.

3.4 Fingering Event Detection

From the above steps, four fingertip positions T_j and four string positions $[P_j, P_j']$, $j = 1, ..., 4$, are obtained for each frame. Based on fingertip positions and string positions, the fingering events, i.e., press, release and fingertip position on the string (string number and distance from the pressing point to the nut), are computed as follows. For each T_j, the string $[P_{j'}, P_{j'}']$ with the smallest point-to-line distance is searched. If the point-to-line distance is smaller than a threshold δ_s, that means the fingertip is pressing the string j'. On the other hand, if no string is found with a distance smaller than δ_s, that means the fingertip is not pressing any string. The press event is detected at the current frame if the fingertip is not pressing any string in the previous frame and is pressing some string in the current frame. Conversely, the release event is detected. If a fingertip is pressing a string, the pressing point distance D_j from T_j to the nut $P_{j'}$ is further calculated for this fingertip.

3.5 Automatic Note Inference

After detected, a fingering event can be automatically translated into a played note. In particular, the press event corresponds to onset, and the release event corresponds to offset.

According to the physics of the vibrating string, the vibrating frequency f is related to the vibrating length L, the tension T and the linear mass of the string U as follows:

$$f = \frac{1}{2L}\sqrt{\frac{T}{U}} \ . \tag{5}$$

As U and T are fixed, L determines the vibrating frequency during violin playing. For a tuned violin in open string case, the vibrating frequencies, $f_{0j}, j = 1$ to 4, of strings E, A, D and G are 659 Hz (E5), 440 Hz (A4), 394 Hz (D4), and 196 Hz (G3), respectively. The distance from the nut to the bridge of ordinary full scale violins L_0 is about 330 mm [12].

When the violin is played, different pitches are produced by pressing different points along the strings (Figure 2). Given the string number j' that is pressed and the distance D_j from the pressing point to the nut, the frequency f_j or

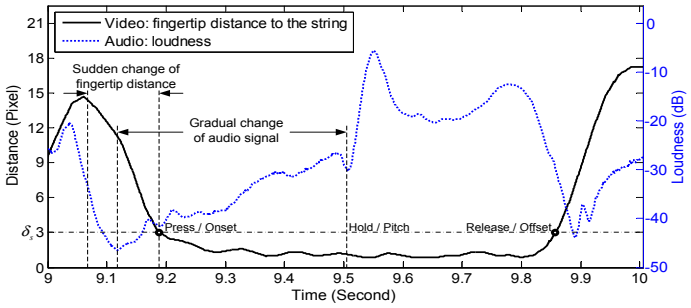

Figure 3: Comparison between audio and video features for music transcription.

pitch of the played note can be computed based on Eq. 5 as:

$$f_j = f_{0j}\frac{L_0}{L_0 - D_j} \ . \tag{6}$$

Figure 4 shows four sample results of the automatic note inference. In Figure 4(a), the fingers start to press string A and D to produce the pitch D5 and $G^{\#}4$. The time instance of this frame corresponds to the onset of the current note and offset of the previous note.

4. SYSTEM EVALUATION

Multiple finger tracking, string detection, and automatic note inference are evaluated using the captured violin fingering video with resolution 720×432, frame rate 25 fps, a total of 5649 frames in 225 seconds, and 504 played notes.

4.1 Evaluation of Multiple finger Tracking

1000 training samples were used to train the joint finger model dynamics. The multiple finger tracking algorithm successfully tracked 20876 of all the 22596 (5649×4) finger contours in 5649 test frames, i.e., a tracking correctness of 92.4%. Some successful test results are shown in Figure 4.

4.2 Evaluation of String Detection

String detection for each string was regarded as successful when the starting and ending points (the nut and the bridge) of the string are correctly detected (Figure 4).

In the experiments, 21289 strings of all 22596 (5649×4) strings in 5649 test frames were correctly detected, i.e., a correctness of 94.2%. In the failure cases, string E detection fails most often, with a total of 786 failures, since it is the thinnest and farthest string in the captured view. The total numbers of failures of strings A, D and G are 125, 165 and 231, respectively.

4.3 Evaluation of Automatic Note Inference

To justify our argument that difficulty for onset/offset detection from gradual change of audio signal can be tackled by extracting press and release events from fingering video, we show a representative case of the gradual change of audio signal and sudden change of fingertip distance in Figure 3. As can be seen, around the onset timing, the change duration of fingertip distance is one third of the gradual change duration of audio signal. Therefore, release event from video provides three times higher accuracy at pinpointing the onset timing. This observation is valid for offset as well.

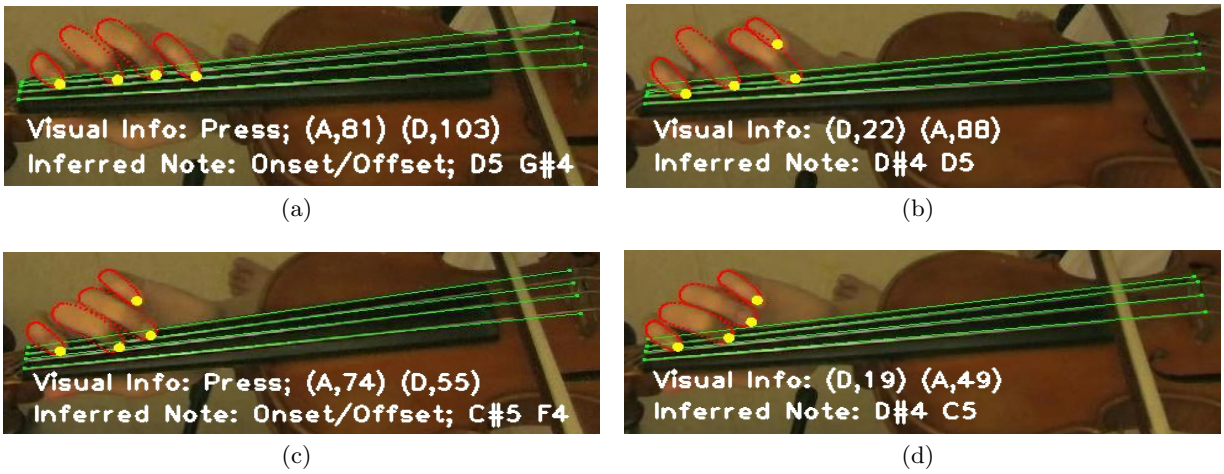

Figure 4: Sample results of multiple finger tracking, string detection, and automatic note inference.

To evaluate the accuracy of automatic note inference, we compared the inferred notes with human annotated notes. If the onset, offset and pitches of an inferred note match the corresponding elements of the human annotated one, we considered it as a full-match. If they match in only one or two elements of the onset, offset and pitches, we considered it as a partial-match. If no element matches, it was considered as a mismatch. Among the 504 played notes, there were 75 full-matches, 328 partial-matches, and 101 mismatches, i.e., 14.9% full-matches and 65.1% partial-matches.

It is worth noting the relatively low accuracy of automatic note inference compared with the high correctness of multiple finger tracking and string detection. One important reason is that in 2-D fingering video, if a fingertip is above a string but not pressing the string, the algorithm will misjudge the fingertip as pressing the string. This drawback results in additional pitches during automatic note inference. This shortcoming can be overcome by employing an additional camera to capture 3-D finger information.

Despite the low accuracy of automatic note inference from video-only approach, the visual information in full-matches and partial-matches can be fused with audio-only data. From preliminary experiments, we found that the visual and audio data are indeed complementary. The results show that the fusion of audio and visual data improves the system performance compared to the audio-only method. This shows that visual analysis of violin fingering has great potential in assisting audio-only music transcription, thus paving the way for an audio-visual violin transcription system. Due to space limitation of this paper, details about audio-visual data fusion will be discussed in other publications.

5. CONCLUSIONS

By exploring the high correlation between the violin playing technique (fingering) and the played acoustic notes, with the first attempt we investigated the visual analysis of violin fingering to compensate for the difficulties of audio-only violin transcription. To achieve this, we proposed a robust multiple finger tracking algorithm and string detection method that detect finger events (press, release and fingertip positions), which are automatically translated into played notes (onset, offset and pitches) based on the physics of the violin. Experimental results have shown high correctness of multiple finger tracking and string detection, thus paving the way for an improved audio-visual violin transcription system.

6. REFERENCES

[1] Oshima C., Nishimoto K., Suzuki M., Family ensemble: a collaborative musical edutainment system for children and parents. *ACM Multimedia,* pp:556-563, 2004.

[2] Yin J., Wang Y. and Hsu D., Digital Violin Tutor: An Integrated System for Beginning Violin Learners, *ACM Multimedia Conf.*, 2005.

[3] Collins N., A comparison of sound onset detection algorithms with emphasis on psycho-acoustically motivated detection functions, *Journal of the Audio Engineering Society,* 2005.

[4] Klapuri A., Automatic music transcription as we know it today, *Journal of New Music Research,* 33(3), pp:269-282, 2004.

[5] Hardenberg C. and Brard F., Bare-hand human computer interaction. *Proc. of Perceptual User Interfaces,* pp:1-8, 2001.

[6] Wu Y., Lin J., Huang T. S., Analyzing and capturing articulated hand motion in image sequences. *IEEE Trans. Pattern Anal. Mach. Intell,* 27(12), pp:1910-1922, 2005.

[7] Burns A. and Wanderley, Visual methods for the retrieval of guitarist fingering. *Proc. of Conf. on New Interfaces For Musical Expression,* pp:196-199, 2006.

[8] Isard M. and Blake A., Contour tracking by stochastic propagation of conditional density, *Proc. European Conf. on Computer Vision,* v1, pp:343-356, 1996.

[9] Blake A. and Isard M., *Active contours,* Springer, 1998.

[10] Hough. P. V. C., *Method and means for recognizing complex patterns,* U.S. Patent, 3.069.654, 1962.

[11] Molteno T. C. A., Tufillaro N. B., An experimental investigation into the dynamics of a string, *American Journal of Physics,* 72(9), pp:1157-1169, 2004.

[12] Bachmann A., *An encyclopedia of the violin,* Da Capo Press, 2005.