iDVT: An Interactive Digital Violin Tutoring System Based on Audio-Visual Fusion

Huanhuan Lu, Bingjun Zhang, Ye Wang and Wee Kheng Leow
School of Computing, National University of Singapore
{luhuan, bingjun, wangye, leowwk}@comp.nus.edu.sg

ABSTRACT

iDVT (interactive Digital Violin Tutor) is a violin learning system exploiting physical and virtual resources and interactivity. It aims at providing the user with new effective learning experience. This demonstration paper briefly describes the structure of the system and the underlying audio-visual processing techniques employed in the system.

Categories and Subject Descriptors

H.5.5 [**Sound and Music Computing**]: Signal analysis, synthesis, and processing; Systems; I.4.8 [**Scene Analysis**]: Motion; Tracking; Sensor fusion

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Music Transcription, Onset Detection, Hand Tracking, Fingering Analysis, Multimodal Fusion

1. INTRODUCTION

Computer-assisted musical instrumental tutoring(CAMIT) is catching eyes of many researchers and musicologist in recent years. Projects providing general instructions are of special interest since they are intended for self-learning and practice, the most frequent scenario encountered by learners in instrument learning. Our system iDVT has similar goals and concerns, but specializes in assisting violin learning, an area less explored in CAMIT previously.

Computer-assisted violin tutoring system requires accurate violin transcription. For pitched non-percussive (PNP) sounds such as from the violin, onset detection is more difficult than pitch estimation. This problem becomes more severe considering the inferior household acoustic environment, where the system is most likely used. State-of-the-art audio-only onset detection approaches in PNP sounds reveal poor performance[2].

To address this problem, our system incorporates video processing to complement the insufficiency of current audio-only method. In the following sections, a brief description of the system and the techniques will be presented.

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.
ACM 978-1-60558-303-7/08/10.

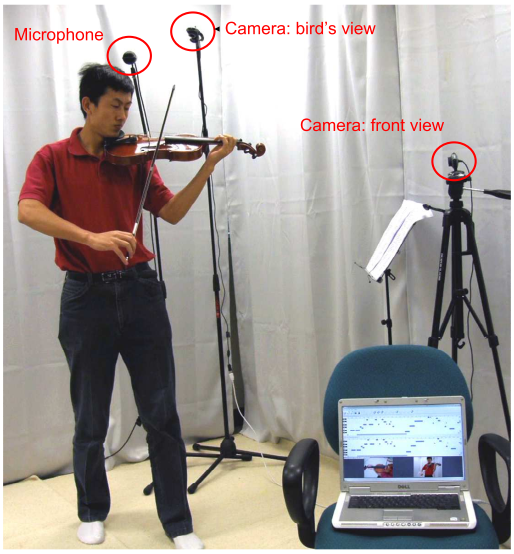

Figure 1: Hardware setting of the system.

2. SYSTEM OVERVIEW

The hardware setting and technical structure of iDVT system are shown in Fig.1 and Fig.2, respectively. iDVT system is used when the learner practices a violin piece following a reference notation. The system has two ordinary webcams and one microphone as peripherals, recording the audio of the playing as well as the videos from the front view(focusing on the bowing) and bird's eye view(focusing on the fingering) of the learner. After the whole recording has completed, the audio and video processing units of the system extract indicative features of onsets (detection functions) from the above three inputs respectively. Subsequently, features derived from audio and videos processing are fused together to obtain a more accurate onset detection result than state-of-the-art audio-only processing. After the onset detection, pitch estimation is conducted at last to produce the MIDI (piano-roll) notation of the played violin music. Through the comparison of the transcribed results and the reference notation(which is prepared beforehand in MIDI), the system manifests every note the violin learner played and indicates which notes are played correctly/wrongly.

The user interface of the iDVT is shown in Fig.3. The first two panels display the piano roll of the reference piece and the transcription result of the user's playing respectively. They are intended for showing how correctly the user played

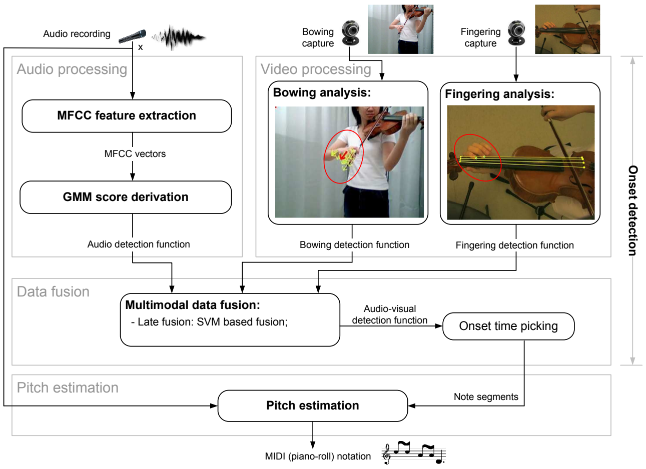

Figure 2: System structure of iDVT.

through comparison between the two. The third panel reflects the user's gesture of playing from two angels and at the same time displays the video processing results. Audio-only processing and audio-visual processing are both supported in the system for performance discretion of the two. All the audio/video raw data and processing results can be evaluated through playback supported by the system.

2.1 Audio Processing

In the audio processing part of the system, a supervised learning approach for onset detection is implemented using Gaussian Mixture Models (GMM) to classify onset and non-onset frames based on Mel-Frequency Cepstral Coefficients (MFCCs) [3] of the input audio. One audio-only onset detection function is derived in this phase.

2.2 Video Processing

In the video capturing the front view of the learner, the right hand conducting bowing is tracked in each frame using Kalman filter framework with measurements obtained by optical flow and a skin color Gaussian model. Through the hand tracking, the bowing direction at any given time is obtained. Moments when the bowing reverses direction are considered as onset times. The bowing detection function can be derived in this phase.

In the video capturing the bird's eye view of the learner, the fingers of left hand are detected using a two step algorithm. Four violin strings are detected first, after which finger positions are searched along each string using the pre-calculated skin-color Gaussian model. Moments when a sudden change of finger positions occurs are considered as onset times. The fingering onset detection function can be derived in this phase.

2.3 Audio-Visual Fusion

In the audio-visual fusion part of the system, the detection functions obtained from audio and video processing are combined to produce an audio-visual detection function more indicative of onsets.

Since the audio and video are recorded simultaneously and time stamped in software level, they are assumed to be synchronized. The three detection functions derived in audio and video processing are interpolated respectively conforming to the same sampling rate and normalized into [0,1]. Subsequently, onsets are obtained after the detection functions are fed into Support Vector Machines (SVM) [1] for decision level fusion.

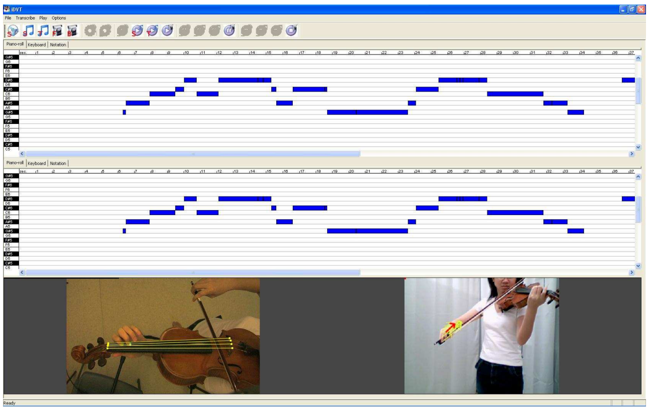

Figure 3: User interface of iDVT.

After onset detection, the violin audio is segmented into individual notes segments and the audio-only pitch estimation is carried out. The pitch estimator evaluated in [4] is employed in our system.

3. CONCLUSION

The system is implemented in C++. It has low hardware requirement which demands only one PC, one microphone, and two ordinary webcams. It is easy to setup since the configuration only involves putting the webcams at the proper positions capturing two views. It is helpful for effective learning since users can see through the system not only how they played but also how well they played. Last but not least, it leverages the synergy of audio and video analysis which provides a well-performed violin music transcription system.

4. ACKNOWLEDGMENTS

This work was supported by Singaporean Ministry of Education grant with the Workfare Bonus Scheme number of R-252-000-267-112.

5. REFERENCES

[1] C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[2] N. Collins. A comparison of sound onset detection algorithms with emphasis on psycho-acoustically motivated detection functions. In *Proceedings of AES118 Convention*, 2005.

[3] B. Logan. Mel Frequency Cepstral Coefficients for Music Modeling. *International Symposium on Music Information Retrieval*, 28, 2000.

[4] Y. Wang, B. Zhang, and O. Schleusing. Educational violin transcription by fusing multimedia streams. *Proceedings of the international workshop on Educational multimedia and multimedia education*, pages 57–66, 2007.