# Toward an Augmented Reality System for Violin Learning Support

Hiroyuki Shiino, François de Sorbier, and Hideo Saito

Graduate School of Science and Technology, Keio University, Yokohama, Japan

{shiino,fdesorbi,saito}@hvrl.ics.keio.ac.jp

**Abstract.** Violin is one of the most beautiful but also one of the most difficult musical instruments for a beginner. This paper presents an on-going work about a new augmented reality system for training how to play violin. We propose to help the players by virtually guiding the movement of the bow and the correct position of their fingers for pressing the strings. Our system also recognizes the musical note played and the correctness of its pitch. The main benefit of our system is that it does not require any specific marker since our real-time solution is based on a depth camera.

**Keywords.** Augmented Reality, Marker-less, Violin pedagogy, depth camera.

## 1 Introduction

Learning how to play violin is very difficult for a novice player. Unlike the guitar, the violin has no frets or marks to help the finger placement. Violinists also have to maintain a good body posture for the bowing movement. Some studies state that a player needs approximately 700 hours to master the basics of violin bowing [1].

Some methods have been introduced to help this learning process. MusicJacket [2] is a wearable system with a vibrotactile feedback that guides the player's movements. However, we consider that wearing such specific device limits the ease of the players since they will not practice under normal conditions. Moreover, this approach does not support the fingering teaching.

Augmented reality technology has the benefit to be non-intrusive and has consequently been applied to musical instrument learning. Motokawa and Saito [3] proposed a guitar support system that displays a computer-generated model of a hand. It helps the player for finger placement and overlays lines where to press the strings. However this kind of approach is using markers [4] added onto the instrument which makes it not robust to occlusions. The limit of markers can be overpassed by using feature point detectors such as SIFT [5]. Although, the texture of the violin is very reflective and uniform which will provide a small number of unstable features that is not adapted for our system. Moreover, feature point detectors are often not robust to illumination changes.

In this on-going research, we proposed a marker-free system using augmented reality for violin pedagogy. It teaches the player where to correctly press the strings on the fingerboard and how to perform the bowing movement by displaying virtual information on a screen. At the same time, our system analyses the musical note played and the correctness of its pitch (frequency of the sound). In this paper, we are aiming at presenting a technical description of our system and not yet focusing on the benefits of its pedagogic side.

We removed the constraint of markers and detectors by including a depth camera that capture the depth information from a scene in real-time. We take advantage of the classic Iterative Closest Point (ICP) algorithm [6] for estimating the pose of the violin based on a pre-reconstructed 3-D model. We also use the human body tracking capability of the depth camera for teaching novice player how to correctly manipulate the bow.

The remainder of the paper is structured as follows: Section 2 briefly gives an overview of our system. The reconstruction of the 3-D model during an offline phase is presented in Section 3. The online phase explaining the tracking, the sound analysis and the display is described in Section 4. Section 5 details how we display the virtual information. Finally, in Section 6 and 7, we present quantitatively our results and our future extensions. Please note that we didn't perform yet any user based studies which will be organized in the future.

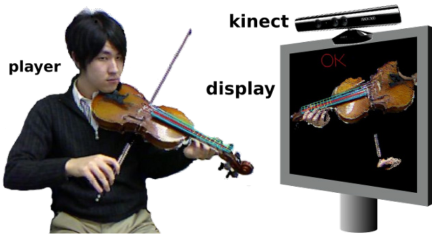

**Fig. 1.** The violinist is captured by a depth camera located over the screen that is a suitable position for both tracking and feedback processing. Virtual advices are displayed on a screen.

## 2 Overview of our Learning System

Our system is based on Kinect[1], a depth camera that captures in real-time a color image and its corresponding depth information. The depth information can easily be converted into a 3-D point cloud using internal parameters of the camera. The violin and the player are extracted from both of these images and analyzed for estimating their pose in the 3-D space. Our learning approach is made of two parts: the first one focus on the finger position while the second try to improve the bowing technique of the player. Virtual information for both approaches is displayed on a screen.

[1] www.kinectforwindows.org

In the first case, it displays the captured violin from a constant viewpoint (even if the player moves then the violin is presented from always the same viewpoint on the screen) with the virtual frets and emphasized strings. In the second case, we display a full view of the player with a virtual skeleton overlaid and containing specific tags located on the bowing arm bones. In the meanwhile, a microphone captures the notes played by the violinist that are analyzed for further virtual advices. An overview of our system is presented in Fig. 1.

## 3 Creation of the Violin Model (Offline)

During the tracking (online phase), we estimate the pose that transforms the observed violin to a pre-computed 3-D model of this violin. This transformation is important because the virtual guides overlaid on the violin are pre-computed in the referential of this 3-D model. This transformation is computed using the ICP [6]. Our experience of the ICP algorithm suggests that a 3-D model defined with too many points will lead to a high computational time. Conversely, a 3-D model described with not enough points will decrease the accuracy of the pose estimation. We decided to separate the 3-D model into several sub-models stored in a database for optimizing the effectiveness of the pose estimation during the tracking phase.

During this offline phase, we capture and segment the violin based on its main color, the depth information and a plane equation. Details about this segmentation will be given in the following section. The sub-models are mainly containing parts from the front face of the violin because it might be the most often observed and important area during the tracking phase. In order to distinguish the sub-models, we describe them with a plane equation computed from the points belonging to front face of the violin. We also use this plane equation to ensure that the sub-models are different enough; each new candidate is then compared with the previous stored ones based on the angle difference between the planes. Finally, the database contains all the sub-models defined with a 3-D point cloud and a plane equation. We set the first-model as the reference for the virtual fingerboard (frets and strings) which is manually added. For this reason, we also store a matrix to remember the transformation from each sub-model to the first one.

## 4 Tracking of the Violin (Online)

To perform the tracking of the violin and the user without the constraint of markers, we are only using the depth information from Kinect. The virtual fingerboard is displayed using again ICP between the pre-computed sub-models and the current captured depth image. To reduce the computational time, we decrease the size of the 3-D point cloud by segmenting the violin based on the color and the depth values. For detecting the human body parts from the depth data, we use an algorithm included in the *OpenNI* SDK. Finally, we analyze the note played in order to give more advice to the novice player. We will describe all those steps in the following sections.

### 4.1 Segmentation of the violin

To reduce the amount of data during the model reconstruction and the tracking, it is better to keep the information only related to the violin's body. Most of the violins have the same brown color, but if we apply only a color segmentation using this information, we might obtain a noisy result with missing information in areas such as the black fingerboard or on the specular parts on the violin's body. We resolved this problem by adding an additional stage after this first color-based segmentation. Thanks to the depth camera, we can get the 3-D points corresponding to the rough color segmentation of the violin. We use these points to compute the violin's front face plane equation minimized with RANSAC. Knowing the common dimensions of the violin, we define a box aligned with the plane and centered at the mean of all the points belonging to the computed plane. All the 3-D points inside of this box are finally registered. Some visual results of our segmentation are presented in Fig. 2.

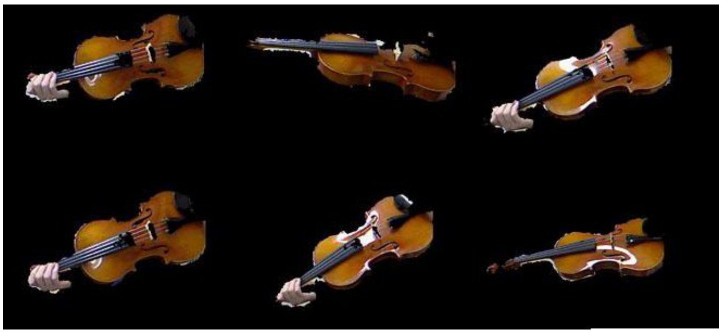

**Fig. 2.** Results of the segmentation. Even specular and occluded parts are correctly segmented.

### 4.2 Tracking of the violin

Our violin's tracking is based on the ICP algorithm applied between the segmented 3-D point cloud of the violin and one of the sub-model stored in the database. This latest is selected by searching for the sub-model with the most similar plane equation.

ICP algorithm results in a rotation matrix and a translation vector that describes the transformation between the captured violin and a sub-model from the database. Since we also know the transformation between the first sub-model (defining the virtual fingerboard), and the other sub-models, we can display the current violin's point cloud in the same referential than the first sub-model. This approach ensures that the virtual information displayed on the screen will always be watch from the same viewpoint even if the player is moving the violin.

### 4.3 User tracking

We propose to advice the novice violinists about the movements of their bow by comparing their gesture with the one from an accomplished player.

Our approach uses the skeleton tracking [7] included in *OpenNI*[2] to capture the movements from both the novice and the experimented players. It detects and tracks in real-time the different parts of the body and deduces from it a skeleton (joints and bones) defined in the 3-D space. The “skilled 3-D skeleton” movements are captured beforehand and replayed during the learning stage. However, the skeletons may not directly match since the novice and the skilled players probably do not have the same body morphology. Our solution is to align the skilled skeleton to the novice’s one by orienting and scaling the axis of shoulders. In that case, the shoulder of the bowing arm will correspond (position and orientation) for both of the skeletons. Finally, we scale the shoulder-elbow and elbow-hand bones from the skilled skeleton to match the size of the novice bones.

### 4.4 Sound analysis

By visualizing the virtual frets and strings, the player can understand where to press to play the violin. However, it remains difficult to recognize if the note played was correct or not. To advise violinists about the correctness of the sound played, we use a spectrum analyzer[3] based on a wavelet transformation to analyze the violin’s sound and to evaluate the accuracy of the pitch in cent unit.

We propose three approaches to select the reference note used for the comparison. In the first one, the system randomly selects a note that the violinist needs to play back. The second one asks the player to select a scale. The system will then ask for the notes on this scale. In the last approach, the player plays the note of his choice that the system recognizes based on the pitch. When the result is displayed, the player can then check if the note played was correct or not.

## 5 Augmented reality based learning support

### 5.1 Bowing support

The players start this learning stage by selecting the string on which they want to practice. Then they need to follow the movements of the skilled violinist that we previously recorded. The parts of the bowing arm (shoulder, elbow and hand) have been emphasized with big dots. The dots from the skilled movement are colored in red while the dots from the novice player are in white. Fig. 3 shows a view of our bowing support system.

We compare the position of the player’s hand and elbow with the one from the skilled skeleton in the 3-D. If the distance is correct then an “OK” mark is displayed. Otherwise a “NG” mark is displayed. Shoulders are not considered since they are supposed to be at the same position for both skeletons. By persevering at maintaining the “OK” position, the player might be able to improve his skills when using the bow.

[2] http://www.openni.org

[3] http://www.fmod.org/

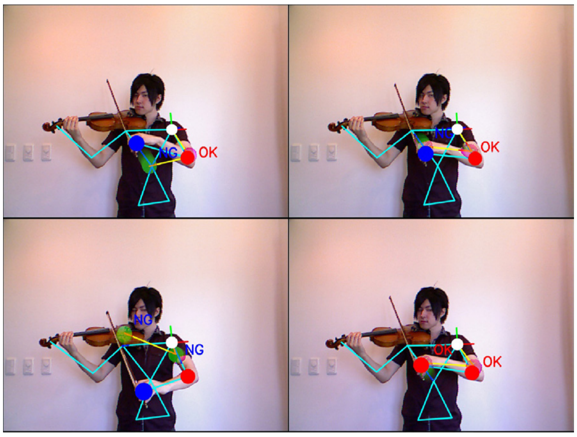

**Fig. 3.** The bowing support emphasized the elbow and the hang of the bowing arm. If the position differs from the pre-recorded movement then a message is displayed.

## 5.2 Displaying the way of playing scale

Our proposed system can teach where to place the finger on the neck of the violin by adding virtual frets and emphasizing the strings. The string and the fret that the violinist needs to press are displayed using respectively a red line and a red dot. Fig. 4 presents the virtual information overlaid onto the violin.

We decided to display always the same viewpoint of the violin to the player by transforming the segmented violin into the first model view. This should allow the user to easily find the useful information on the screen since the virtual frets and strings will always be located at the same position.

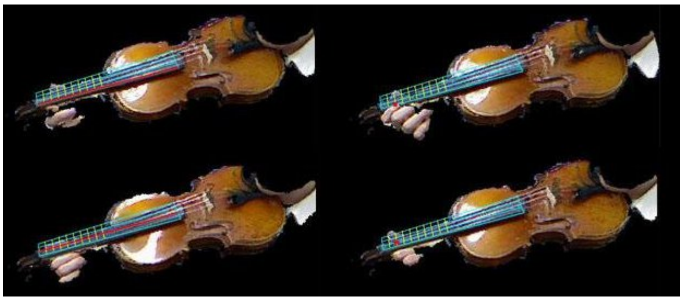

**Fig. 4.** Left side: The string that the player needs to press is in red. Right side: The fret that has to be pressed is marked with a red dot.

## 5.3 Displaying the sound analysis result

If we consider that the player is correctly pressing the strings with the fingers then the goal of the sound analysis is also to verify that the position of the bow on the strings is correct. One example of our learning stage using the pitch's accuracy of the note played is depicted in Fig. 5. When the user plays at the correct pitch, an "OK" mark is displayed. If the pitch is too low or too high, then the "Low" or "High" marks appear. In this latest case, a green arrow is also displayed to indicate to the player the direction where the bow has to be moved to get the correct pitch.

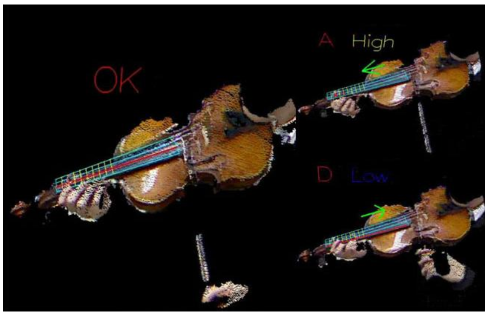

**Fig. 5.** Information is displayed to advise the position of the bow on the strings depending on the correctness of the pitch.

## 6 Results

Experiments were performed on an "Intel Core2 DUO 2.80GHz" PC. We measured an average computational time of 21ms (~45 frames per second) that is suitable for a real-time rendering. For this experiment, we first evaluated the accuracy of our tracking approach based on ICP. We compared it with the AR-Toolkit marker tracking while trying to avoid occlusions of the markers. We added four markers on the body of the violin and pre-computed the sub-models based with it. During the online phase, we compute the rigid transformation between the first sub-model and the segmented violin using our approach and using the marker-based approach. Considering the marker-based transformation as the ground truth, we had the results presented in Table 1. Even if our results seem a little bit less accurate, our approach has still the benefit to be robust against occlusions.

**Table 1.** Evaluation of our tracking compared to the ground truth. It shows the rigid transformation matrix decomposed in three rotations an one translation.

| | **Rx(deg)** | **Ry(deg)** | **Rz(deg)** | **T(mm)** |
|---|---|---|---|---|
| Minimum error | 0.12 | 0.25 | 0.20 | 0.22 |
| Maximum error | 13.29 | 8.27 | 7.89 | 32.1 |
| Average error | 3.07 | 2.69 | 2.78 | 7.20 |

We also evaluated the accuracy of the virtual frets' position. Each fret has a corresponding pitch, so by pressing the strings we expect to obtain a similar pitch. For this experiment, we measured the correctness of the pitch when a skilled player (to ensure a correct manipulation of the bow) was using the virtual frets. Table 2 presents the results of this experiment for each fret where a difference of pitch closes to zero means that the accuracy is good. These results show that the position of the frets is almost correct.

**Table 2.** Difference of pitch (in cent unit)

| **Fret number** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **Average** |
|---|---|---|---|---|---|---|---|---|---|---|
| Difference of pitch | 11.1 | 14.1 | 12.0 | 12.4 | 13.4 | 15.8 | 12.8 | 13.9 | 19.2 | 13.8 |

## 7 Conclusions

We have presented the technical part of our on-going work on a marker-free augmented reality system for assisting the novice violinists during their learning. Thanks to a depth camera, we are able to advise the player on his fingering and bowing techniques by displaying virtual information on a screen.

Our next step is to perform a user based study with novice and skilled players to confirm our choices. Finally, we are also working on a see-through HMD version of our system for a better view of the virtual information directly on the violin.

## References

1. J. Konczak, H. vander Velden, L. Jaeger. Learning to play the violin: motor control by freezing, not freeing degrees of freedom by freezing. Journal of motor behavior, 41(3):243-252, 2009.
2. J. van der Linden, E. Schoonderwaldt, J. Bird, R. Johnson. MusicJacket - Combining motion capture and vibrotactile feedback to teach violin bowing. IEEE Transactions on Instrumentation and Measurements, Special issue on Haptic, Audio and Visual Environments for Games, 2009.
3. Y. Motokawa, H. Saito. Support system for guitar playing using augmented reality display. In Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality, 243-244, 2006.
4. H. Kato, M. Billinghurst. Marker tracking and HMD calibration for a video-based augmented reality conferencing system. In Proceedings of the 2nd International Workshop on Augmented Reality, 1999.
5. D. Lowe. Object recognition from local scale-invariant features. Proceedings of the International Conference on Computer Vision, 2: 1150–1157, 1999.
6. Z. Zhang. Iterative point matching for registration of freeform curves and surfaces. International Journal of Computer Vision, 13(2):119-152, 1994.
7. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio. Real-time human pose recognition in parts from single depth images. In Proc. of IEEE CVPR, 2(7), 2011.