Hand-by-Hand Mentor: An AR based Training System for Piano Performance

Ruoxi Guo[1*] Jiahao Cui[1†] Wanru Zhao[1‡] Shuai Li[1,2§] Aimin Hao[1,2¶]

State Key Laboratory of Virtual Reality Technology and System, Beihang University[1]
Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University[2]

Abstract

Multimedia instrument training has gained great momentum benefiting from augmented and/or virtual reality (AR/VR) technologies. We present an AR-based individual training system for piano performance that uses only MIDI data as input. Based on fingerings decided by a pre-trained Hidden Markov Model (HMM), the system employs musical prior knowledge to generate natural-looking 3D animation of hand motion automatically. The generated virtual hand demonstrations are rendered in head-mounted displays and registered with a piano roll. Two user studies conducted by us show that the system requires relatively less cognitive load and may increase learning efficiency and quality.

Index Terms: Human-centered computing—Virtual Reality—; Human computer interaction—User evaluation

1 Introduction

To facilitate both piano teaching and learning process, a variety of multimedia teaching patterns have been proposed in recent years, including (but not limited to) prerecorded teaching video, 2D performance animation [2,5], and interactive notation projection [1,3]. However, the majority of the previous teaching patterns concentrate on presenting rigid instructions to direct users to strike piano keys with correct fingers yet fail to depict intuitive or immersive finger movements.

In this paper, we present Hand-by-Hand Mentor, an individual training system based on augmented reality (AR) for piano performance. Our training system, which can provide virtual performance animation on a (physical) piano, overcomes three main challenges in teaching and learning the piano, which is reducing cognitive load in sight-reading, generating 3D performance animation automatically, and presenting visual feedback in an immersive environment.

As shown in Fig. 1, given a Musical Instrument Digital Interface (MIDI: https://en.wikipedia.org/wiki/MIDI/) file as input, our system generates 3D performance animations for both hands and these animations are rendered in a head-mounted display (HMD), aligned with a real piano. The 3D performance animations can well contribute to realistic and appealing demonstrations/instructions in AR environments and are easy to follow.

We evaluate the AR-based piano training system based on two user studies, comparing it to two traditional training patterns (video training and notation projection training). Results show that our system provides more realistic and appealing performance demonstrations, and may increase learning efficiency and quality.

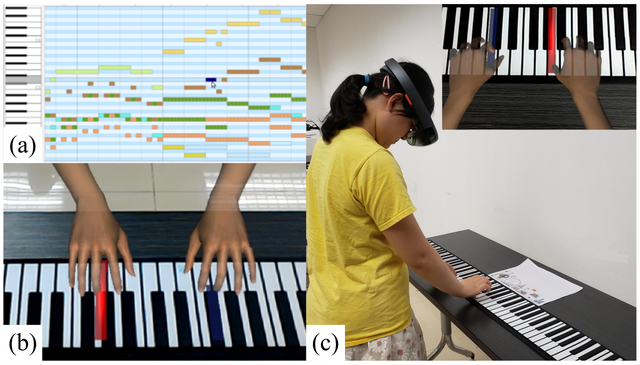

Figure 1: (a) The input MIDI file contains the time sequences of different keys. (b) Based on the MIDI file, our system generates 3D performance animations for both hands. (c) Performance animations are rendered in a HMD, aligned with a real piano, to provide users with real-time visual guidance and feedback.

2 System Design and Implementation

2.1 Automatic Performance Programming

In this study, the state of fingers and the state of piano keys are encoded in the topology of the HMM. The feature space is composed of the 10 fingers (inputs) and the 88 keys (outputs). The key features represent the HMM's observations and the fingers control the HMM's hidden variables.

We label the fingers from thumb to pinky as 1-5 for the right hand and 6-10 for the left hand. We define the probability of obtaining a certain key K_i as the probability that we will play K_i with finger F_i after we play the key K_{i-1} with the finger F_{i-1}, which is denoted as $P(K_i|(K_{i-1},F_i,F_{i-1}))$. We assumed that the current state only depends on the previous state.

The feature space is composed by known variables and unknown variables that we want to predict. The prediction consists of first, finding the most likely path in our HMM given an observed sequence of keys (variables) and second, computing a Gaussian Mixture Regression (i.e. predict variables) for each state in the predicted path. Performed by means of the Viterbi algorithm [4], a technique that finds the state sequence or path $f = F_1, F_2, ...F_{10}$ that most likely generated the complete sequence of observations $k = K_1, K_2, ...K_{88}$, that is finding the path such that,

$$\begin{aligned} f &= \arg\max_f (P(f|k)) \\ &= \arg\max_f \left\{ p(F_0)p(K_0|F_0)\prod_{i=0}^{t} p(F_i|F_{i-1})p(K_i|F_i)) \right\} \end{aligned} \tag{1}$$

where $p(F_0)$ is the prior probability of state F_0. This criterion is hence of global optimality.

[*]e-mail: rebeccag@buaa.edu.cn, co-first author
[†]e-mail: cuijh50892@126.com, co-first author
[‡]e-mail: zhaowrenee@gmail.com
[§]e-mail: lishuai@buaa.edu.cn
[¶]e-mail: ham@buaa.edu.cn

Table 1: The actions number for each finger in one hand

Finger	1/6	2/7	3/8	4/9	5/10	wrist	sum
numbers	2	5	5	4	3	2	22

2.2 Piano Fingerings and Skills

Although we can automatically decide which finger to be used for a particular key in a key sequence, the transition of different fingers, which is important for performance, is still a challenge for novices. Piano performance requires complex pose sequences for both hands involving 32 joints. The high dimension of the solution domain makes it difficult to find an optimal solution for joint parameters using an optimization algorithm, which leads to unnatural or impractical hand motions. Therefore, we refine the solution domain by modeling piano fingerings and skills including natural fingering, crossing fingering, and extending/contracting fingering.

After building an animation hierarchy consists of an arm, wrist, and fingers from top to bottom, for each hand we need an animation database containing 22 animations of fingers and wrist animations to cover the commonly seen fingerings in our model. Four transition animations are needed to realize the finger crossing. Table 1 shows the number of animations needed for each finger and wrist in one hand. To further realize the finger crossing of both hands, we also added four transition animations.

2.3 Performance Animation Generation

Based on the animation database, we propose a novel algorithm to refine the solution domain and generate 3D performance animations automatically by combining musical prior knowledge and optimization methods.

Unlike most previous methods deciding the transition by calculating cost, we simplify the decision progress by prioritized transition strategy.

(1) Using displacement to complete the note transition. The hand moves directly to the desired position, and the hand posture remains unchanged.

(2) Using only finger-crossing to complete the transition between fingers and fingers.

(3) Using only finger-extension or contraction to complete the note transition.

(4) Mixing mechanisms, such as using displacement and finger-extension simultaneously.

In view of piano-playing regulations and our hand structure, finger-crossing and finger extension/contraction can't happen at the same time. In our design, therefore, cross-finger and displacement won't be used at the same time. If the cross-finger can not complete the transition without hand displacement, we would use hand displacement directly(category (1)). Thus, we only consider combining (1) and (3) in this mechanism.

3 Evaluation Result

We conducted two user studies to compare Hand-by-Hand Mentor to two traditional training patterns, which are AR display notation projection training and screen display video training. Through cognitive load, we can see whether a training pattern is suitable for novices.

Fig. 2 shows the cognitive load results. Hand-by-Hand Mentor induces lower intrinsic and extrinsic load than video training and higher germane load than notation projection training and video training.

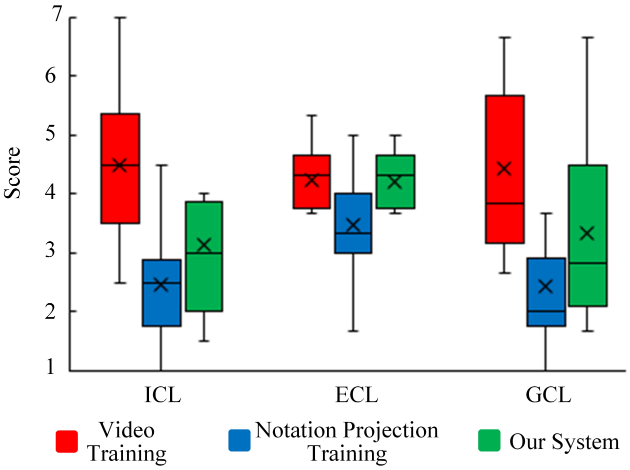

Figure 2: Cognitive load results.

In sum, our system which can also demonstrate 3D hand motion guidance requires far less cognitive load compared with video training. Despite requiring a little more cognitive load than AR display notation projection, our method provides training of higher quality which contains human-like hand posture guidance and feedback. The system supports a better user experience with significantly less cognitive load and increases learning efficiency and quality, especially for novices who heavily rely on visual hand motion guidance and feedback.

4 Conclusion

In this paper, we have presented Hand-by-Hand Mentor, an AR-based individual training system. There are two main contributions of the system we presented: (1) a framework which directly takes MIDI files as input and provides virtual performance instructions and feedback on real pianos; (2) a novel performance animation generation algorithm which combines musical prior knowledge with optimization theory and finally contributes to natural and comfortable hand motions. The results of our user studies showed that the learning process with Hand-by-Hand Mentor may help beginners gain higher efficiency and better performance in music playing.

References

[1] C. A. T. Fernandez, P. Paliyawan, C. C. Yin, and R. Thawonmas. Piano learning application with feedback provided by an ar virtual character. In *2016 IEEE 5th Global Conference on Consumer Electronics*, pages 1–2. IEEE, 2016.

[2] N. Kugimoto, R. Miyazono, K. Omori, T. Fujimura, S. Furuya, H. Katayose, H. Miwa, and N. Nagata. Cg animation for piano performance. In *SIGGRAPH'09: Posters*, pages 1–1. 2009.

[3] K. Rogers, A. Röhlig, M. Weing, J. Gugenheimer, B. Könings, M. Klepsch, F. Schaub, E. Rukzio, T. Seufert, and M. Weber. Piano: Faster piano learning with interactive projection. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces*, pages 149–158, 2014.

[4] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

[5] Y. Zhu, A. S. Ramakrishnan, B. Hamann, and M. Neff. A system for automatic animation of piano performances. *Computer Animation and Virtual Worlds*, 24(5):445–457, 2013.