Multi-Modal Deep Learning-Based Violin Bowing Action Recognition

Bao-Yun Liu[1], Yi-Hsin Jen[2,3], Shih-Wei Sun[4], Li Su[2], and Pao-Chi Chang[1]
[1] Dept. Communication Engineering, National Central University, Taiwan
[2] Institute of Information Science, Academia Sinica, Taiwan
[3] Dept. Computer Science, National Tsing Hua University, Taiwan
[4] Dept. New Media Art, Taipei National University of the Arts, Taiwan

***Abstract*—In this paper, a deep learning-based violin action recognition is proposed. By fusing the sensing signals from depth camera modality and inertial sensor modalities, violin bowing actions can be recognized by the proposed deep learning scheme. The actions performed by a violinist are captured by a depth camera, and recorded by wearable sensors on the forearm of a violinist. In the proposed system, 3D convolution neural network (3D-CNN) and long short-term memory (LSTM) deep learning algorithms are adopted to generate the action models from depth camera modality and inertial sensor modalities. The features and models obtained from multi-modalities are used to classify different violin bowing actions. A fusion process from different modalities can achieve satisfactory recognition accuracy. In this paper, we generate a violin bowing actions dataset for the preliminary study and the system performance evaluation.**

I. Introduction

Action recognition is becoming a popular research area due to many wearable IoT devices widely used in many applications. Most action recognition systems focus on recognizing gestures with large motions. To name a few, Liu et al. [1] proposed a large scale benchmark for multimodal human action understanding (multi-cameras, depth cameras, and skeletons), and PKU-MMD dataset is provided for further study. In addition, Chen et al. [2] proposed to use a Kinect camera and inertial sensors to record the sensing data of a user for daily life action recognition, and UTD-MHAD dataset is also provided. However, only few researchers [3] paid attention to bowing action recognition for violin performance based on wearable sensors. Therefore, in this paper, focusing on violin bowing actions recognition, we propose a multi-modal deep learning structure for a preliminary study.

II. Proposed Method

In this paper, the deep learning structure of the proposed violin bowing action recognition system is shown in Fig. 1. The system is divided into 3 parts. On the left part of Fig. 1, the depth frames captured from a depth camera are sequentially taken as the input for a 3D CNN [4] deep learning process, and the detail processes are shown by the blocks in Fig. 1. The proposed 3D CNN includes two 3D convolution layers and two max pooling layers. For the depth frames, the adopted 3D CNN is used to analyze the motion information from the successive frames of a bowing action observed in a period of time. By adding 3D convolution layers, the spatiotemporal features can be obtained.

On the other hand, as shown by the middle and the right part of Fig. 1, two identical LSTM deep learning processes are adopted for analyzing the inertial data collected from the wearable sensor, i.e., accelerometer and gyro sensor, worn on a violinist. For the inertial sensing modality, the LSTM can analyze the temporal relationship of a bowing action from a violin performer. In addition to the raw data of the accelerometer and the gyro sensor, the total amount of accelerometer and the total amount of the gyro sensor are also used for the LSTM process to train the bowing action model.

Once the classifiers of the depth frames modality, accelerometer modality, and gyro sensor modality are generated, a decision-level fusion process is used for violin bowing action recognition.

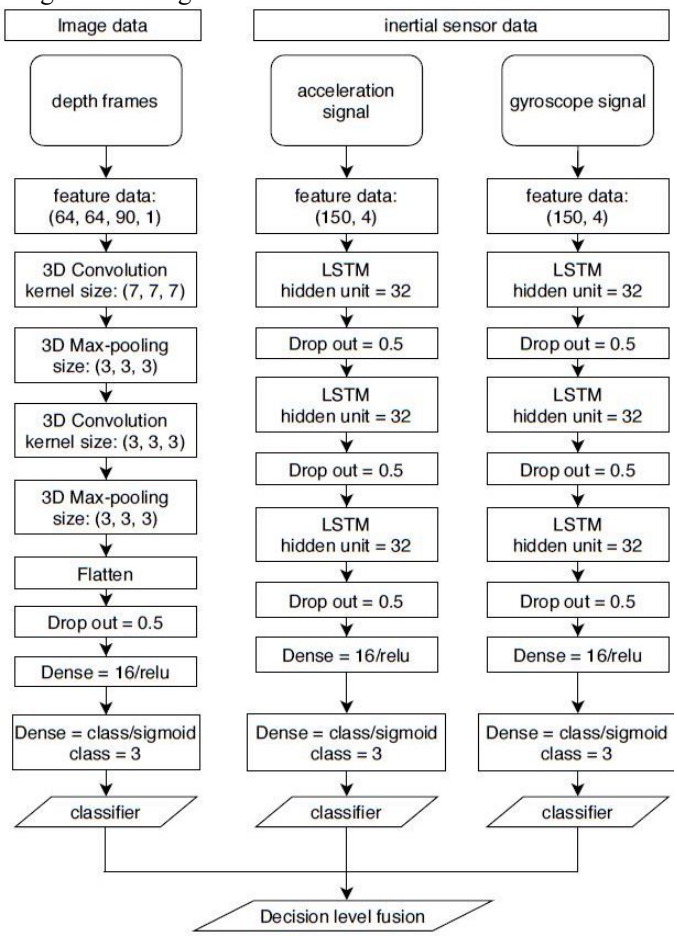


Figure 1. The deep learning structure of the proposed multi-modal violin bowing action recognition system

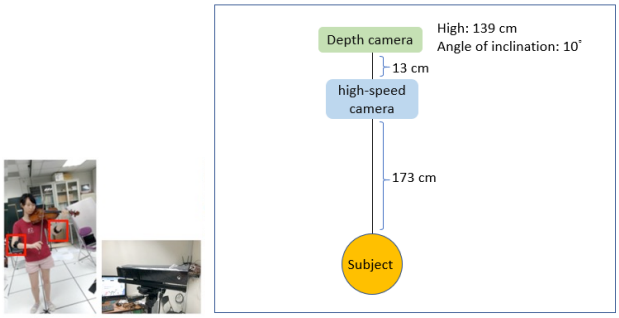
III. Dataset

In order to record the bowing actions with small motions from a violinist, a experimental dataset is generated by cutting from a violin playing for a complete piece of an artwork. A

978-1-7281-7399-3/20/$31.00 ©2020 IEEE

violin playing expert is invited to tag the played bowing actions in different time periods. The sensors used for generating the dataset include two Myo sensors (Fig. 2(a)) and one Kinect camera (Fig. 2(b)), and the geometric relationship of the testing environment is depicted in Fig. 2(c). The sensing data from wearable sensors and the depth frames captured from the Kinect camera are individually recorded from the multiple computers, and the start/end recording instructions are transmitted by TCP sockets from a server computer. Furthermore, a representative recorded reference color frame, depth frame and the corresponding skeletons are shown in Fig. 3. Therefore, a synchronized violin bowing action dataset from multi-modal sensors can be obtained.

In this preliminary study, we focus on recognizing three violin bowing actions: "up bow", "down bow", and "Detache", as shown in Fig. 4. The volunteering violinist are violin major students from the music department, Taipei National University of the Arts. The performers use their left hand to handle a violin and the right hand to hold the bow. "Up bow" and "down bow" are pressing the string on a violin using a finger and move the bow to upward and downward directions, respectively. "Detache" is to operate some pressure on a string to make the bow to be pushed on the string, and play the note symbols independently.



(a) (b) (c)

Figure 2. The recording environment of the proposed violin bowing action recognition: (a) two Myo inertial sensors worn on a violinist, (b) one Kinect camera mounted in front of a violin player, and (c) the geometric relation of the recording environment.

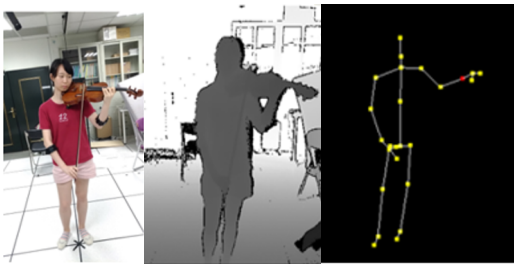


Figure 3. A representative bowing action in color frame, depth frame, and the skeletons.

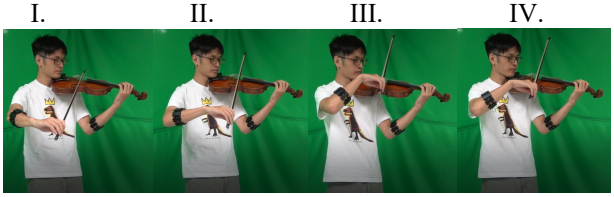


Figure 4. The representative frames of violin bowing actions from a color camera experimental result

IV. EXPERIMENTAL RESULTS

In the experimental results, for each sample, the depth frames are normalized to 50 sample/s, and the inertial sensing data from multiple modalities are resample to 150 sample/s. In this preliminary study, we use the data from two performers to operate 3 different violin bowing actions, and the performer operate the actions for 14 times. We use 80% samples for training and the other 20% samples for testing. The confusion matrix is shown in Table 1. Based on the proposed deep learning structure with a decision level fusion process, the accuracy of "up bow", "down bow", and "Dateche" are ranged from 82.6% to 91.7%. The preliminary study verified that using the depth camera modality and inertial sensor modality can provide a satisfactory violin bowing action recognition capability. However, as shown in Table 1, 17.4% of the "down bow" are falsely recognized as "up bow", and 12% of "Up bow" are falsely recognized as "down bow". The actions with small motion manner are challenging to be recognized in only a very short period of time, not only from the depth frames, but also from the sensing data. In the future, the challenging violin bowing actions deserve more and more researchers to make contributions, and more technologies for a short-term small motions recognition should be developed.

Table 1

True\Prediction	Up bow	Down bow	Detache
Up bow	88%	12%	0%
Down bow	17.4%	82.6%	0%
Detache	5.5%	2.8%	91.7%

Total accuracy: 88.1%

REFERENCE

[1] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A Large Scale Benchmark for Continuous Multi-Modal Human Action Understanding", arXiv:1703.07475 [cs.CV], 2017

[2] https://personal.utdallas.edu/~kehtar/UTD-MHAD.html (Chen et al., IEEE ICIP 2015)

[3] D. Dalmazzo and R. Ramírez, "Bowing Gestures Classification in Violin Performance: A Machine Learning Approach", Frontier in Psychology, 10: 344, March, 2019.

[4] Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 35, No. 1, 221-231, 2012.