applied sciences

MDPI

Article

Audio-Visual Tensor Fusion Network for Piano Player Posture Classification

So-Hyun Park and Young-Ho Park *

Department of IT Engineering, Sookmyung Women's University, Seoul 04310, Korea; shpark@sm.ac.kr
* Correspondence: yhpark@sm.ac.kr

Received: 9 September 2020; Accepted: 28 September 2020; Published: 29 September 2020

Abstract: Playing the piano in the correct position is important because the correct position helps to produce good sound and prevents injuries. Many studies have been conducted in the field of piano playing posture recognition that combines various techniques. Most of these techniques are based on analyzing visual information. However, in the piano education field, it is essential to utilize audio information in addition to visual information due to the deep relationship between posture and sound. In this paper, we propose an audio-visual tensor fusion network (simply, AV-TFN) for piano performance posture classification. Unlike existing studies that used only visual information, the proposed method uses audio information to improve the accuracy in classifying the postures of professional and amateur pianists. For this, we first introduce a dataset called C3Pap (Classic piano performance postures of amateur and professionals) that contains actual piano performance videos in diverse environments. Furthermore, we propose a data structure that represents audio-visual information. The proposed data structure represents audio information on the color scale and visual information on the black and white scale for representing relativeness between them. We call this data structure an audio-visual tensor. Finally, we compare the performance of the proposed method with state-of-the-art approaches: VN (Visual Network), AN (Audio Network), AVN (Audio-Visual Network) with concatenation and attention techniques. The experiment results demonstrate that AV-TFN outperforms existing studies and, thus, can be effectively used in the classification of piano playing postures.

Keywords: piano playing posture; audio-visual tensor fusion; classification

1. Introduction

Studies on classifying playing postures are being carried out in various fields. Research on piano playing posture classification can be used for piano education, playing posture training, and evaluation systems, making these studies important. Moreover, identifying the correct posture during piano performance enables a pianist to protect his body from various Playing Related Musculoskeletal Disorders (PRMDs) [1].

To prevent PRMDs, a variety of methods have been proposed, including the analysis of players' posture through wearable and physical devices. For example, [2] proposed a motion capture system integrated with a data glove that can visualize the skeleton of the pianist's arms and hands. However, most of these devices are expensive and uncomfortable to use. As a result, alternative methods that use visual information are proposed. For example, the authors of [3,4] proposed a method that uses Kinect depth sensors to recognize the pianist's head, shoulder, arm silhouette, elbow, and wrist. The authors of [5] analyzed the differences in finger movements between professional and amateur pianists using statistical analysis methods. The authors of [6] proposed a piano playing posture training method that calculates the skeleton error rate of the teacher and student detected by the Kinect sensors. However, there are certain cases when the posture classification accuracy may degrade because of deteriorating

video quality and rapid hand movement. Thus, in this paper, we propose to utilize audio information in addition to visual information due to the deep relationship between posture and sound.

This study proposes an Audio-Visual Tensor Fusion Network (AV-TFN), the first deep learning-based method for piano playing posture classification method using audio-visual information. The main idea of the AV-TFN is a novel data representation method that can preserve the identity of each piece of audio-visual information. More precisely, the contributions of this paper are as follows.

- We first propose a dataset called C3Pap (Classic Piano Performance Postures of Amateur and Professionals). For this study, we collected videos of both professional and amateur pianists from the YouTube platform. The main advantage of the proposed dataset is that we collected actual performance videos in diverse environments, unlike existing datasets that do not consider various situations of the pianist.
- Second, this study proposes an audio-visual fusion method that represents the audio-visual information as a data structure. The proposed audio-visual fusion method expresses audio information on the color scale and visual information on the black and white scale. It mixes the two colors to represent one data structure. As such, it is a data representation method that can retain audio-visual identity in one data structure and represent relativeness between audio and video information for piano playing posture classification.
- Third, this study demonstrates the superiority of the proposed AV-TFN method through comparisons of the performances of the visual network (VN) [7], audio network (AN), and audio-visual network (AVN) with concatenation (AVN-Concat) [8] and attention (AVN-Atten) [9] techniques. The experiment results show that AV-TFN significantly improves F1 score compared with AN, VN, AVN-Concat, and AVN-Atten methods, while also achieving speeds similar to that of the fast VN method.

This study is organized as follows. Section 2 discusses related studies. Section 3 introduces the proposed AV-TFN. Section 4 presents the experiments and analyzes the results. Finally, Section 5 concludes the paper and highlights future research.

2. Related Work

In this section, we first discuss related studies on player posture in the piano education field and describe their limitations.

2.1. Piano Playing Posture Classification Methods

There have been numerous studies of piano playing posture classification in the past. Several studies use wearable devices distributed over the human's body and collect the motion data from them. For example, [2] proposed a motion capture system integrated with a data glove that can visualize the skeleton of the pianist's arms and hands. The authors of [10] proposed a piano playing posture training system that uses VICON MX 3D Motion Capture System. This system overlays the teacher and student's postures to help the student play with the correct posture. However, most of these wearable devices are expensive and uncomfortable to use. As a result, alternative methods that use inexpensive devices, such as Kinect, are proposed. The authors of [3,4] proposed a method that uses Kinect depth sensors to recognize the pianist's head, shoulder, arm silhouette, elbow, and wrist. The authors of [11,12] investigated the classification of piano hand posture using Kinect depth sensors. The authors of [6] proposed a piano playing posture training method that calculates the skeleton error rate of the teacher and student detected by the Kinect sensors. Using statistical analysis methods, ref. [5] analyzed the differences in finger movements between professional and amateur pianists. The authors of [13] used principal component analysis (PCA), a statistical procedure, to analyze pianists' posture and classify them into various levels. The authors of [7] performed a feasibility test that recommends a suitable deep learning algorithm to determine the correct posture for a pianist. Though this is not a

study on posture classification, [14] proposed a Long Short-Term Memory (LSTM)-based model for predicting playing posture from sound while emphasizing the relationship between posture and sound.

Recall from Section 1 that there are certain cases in which the accuracy of posture classification may degrade due to insufficient visual information. For example, Figure 1a shows a case when the skeletal recognition rate decreases due to deteriorating video quality. Moreover, the skeleton recognition rate and may decrease when distortion occurs, such as a hand covering the piano or when the camera zooms in and out (refer to Figure 1b). On the other hand, the accuracy of posture classification may degrade due to the characteristics of the piano domain. For example, when a staccato is played, the hand momentarily rises upward from the reaction. When the wrist quickly rises, the skeleton recognition rate decreases, and the posture classification accuracy decreases, as shown in Figure 1c. Furthermore, because playing a rotation requires alternating between two notes with a large leap, left and right rotation of the wrist occurs, thus decreasing the skeleton recognition rate (refer to Figure 1d). Besides, sudden movements, such as suddenly crouching or lifting the shoulders or elbows, may also deteriorate the accuracy of posture classification.

(**a**) (**b**) (**c**) (**d**)

Figure 1. Limitations of existing methods. The posture classification accuracy may degrade when the skeletal recognition rate decreases due to (**a**) deteriorating video quality; (**b**) hand covering the piano; (**c**) rapid hand movement; (**d**) hand rotation.

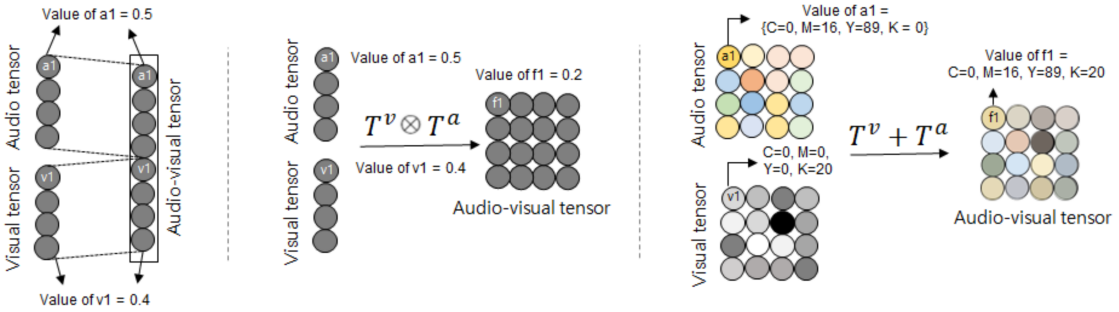
The main feature of the proposed method is to utilize audio information to compensate for cases in which the accuracy of posture classification deteriorates. As audio information is related to posture [15], it can be used to enhance posture classification accuracy even if the skeleton recognition rate decreases. We use audio information in addition to video data due to the following reasons. First, when using audio data, the patterns of notes played for each technique differ. Therefore, audio data facilitate the classification of piano techniques in cases of insufficient visual information. Second, the sound produced by professional pianists and amateur pianists differ. Specifically, that of professional pianists is more resonant and rounder than that of amateur pianists. Audio data thus can also facilitate the classification of the postures of professional and amateur pianists by technique.

2.2. Data Representation Methods

To the best of our knowledge, there are no methods that use video and audio information to classify postures in the music domain. Thus, we investigated studies on representing audio-visual information to solve various problems in other domains. We can classify these studies into two types. The first type involves extracting features from each data and connecting the feature points in one dimension [8,16]. In contrast, the second type involves extracting a feature vector from each datum and performing a Cartesian product operation [17].

The existing related studies of the first type include the works of Poria et al. [8] and Morency et al. [16], who fused multi-modal data that contain text, acoustic, and visual information for emotion analysis. In [8,16], features of length n were extracted from multi-modal data and combined into a single feature (refer to Figure 2a). The existing related studies of the second type include the work of Zadeh et al. [17], who proposed a tensor fusion network that fuses visual, acoustic, and language information through the Cartesian product operation for emotion analysis. In this approach, three

modal data are firstly expressed as tensors with the same length n. After that, the Cartesian product operation is performed using pairs of these tensors. However, in existing studies [8,16,17], since the feature vector of the data before fusion cannot be known, it is difficult to express relevance. For example, in Figure 2b, the calculated values of audio feature a1 (0.5) and visual feature v1 (0.4) yield to a fusion of f1 = 0.2. With the data fusion result of 0.2, it is unknown whether the feature vector before fusion was (0.5, 0.4) or (0.2, 1.0). In other words, existing data representation methods are limited in that the relationships among data cannot be learned simply by connecting them.



(**a**) (**b**) (**c**)

Figure 2. Comparison of the proposed data representation method with existing ones: (**a**) Simple concatenation (Adapted from [8,16]); (**b**) Cartesian product (Adapted from [17]); (**c**) proposed method (AV-TFN).

To overcome these limitations, we developed a novel color-based data representation method that can express the relationship between audio-visual information by preserving the characteristics of the original data even after fusion. For example, in Figure 2c, the Cyan Magenta Yellow and Key (CMYK) values, obtained by changing the feature vector into a color according to a specific rule, are stored. Here, the result is obtained by mixing the color tone (red) as the audio feature and the brightness (dark) as the visual feature is dark red. Because of the colors are mixed by rules, even after mixing, it is possible to estimate what colors were mixed. This solves the problem of the inability to track feature vectors before fusion in the existing methods and helps the deep learning model encode meaningful information.

3. Proposed Audio-Visual Tensor Fusion Network

Recall from Section 1 that this study proposes an AV-TFN, the first deep learning-based piano playing posture classification method using audio-visual information. Figure 3 demonstrates the overall process of AV-TFN. We first (a) collect the C3Pap dataset with five piano techniques; (b) extract audio features using Mel Frequency Cepstral Coefficient (MFCC), which is a well-known technique used in automatic speech and speaker recognition, and visual features using OpenPose [18], which enables us to extract skeleton information of the human body; (c) produce audio-visual tensor fusion using MFCC and skeleton data; (d) classify audio-visual tensors using proposed AV-TFN. Subsequent sections describe each step of the AV-TFN in detail.

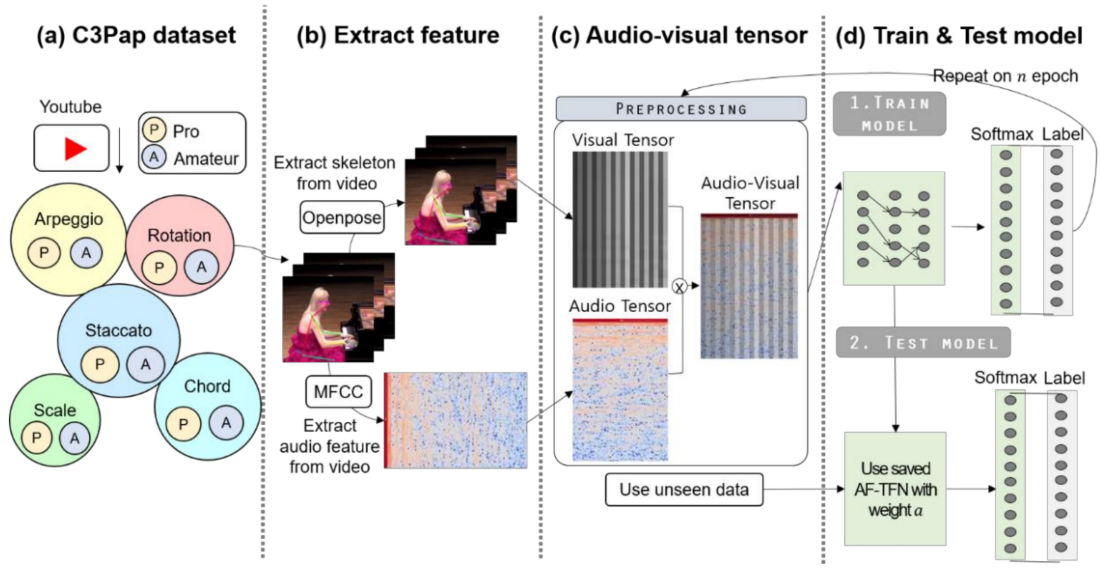


Figure 3. Overall process of AV-TFN.

3.1. Explanation of C3Pap Dataset

We first describe the proposed dataset called C3Pap. C3Pap consists of videos of professional pianists and amateurs playing scales, arpeggios, and chords based on [11,12]. Besides, Sandor, a Hungarian-American pianist introduced technical or physical movement patterns fundamental to playing the piano; accordingly, this dataset added staccato and rotation techniques based on [19]. Details of representative piano playing techniques are shown in Figure 4.

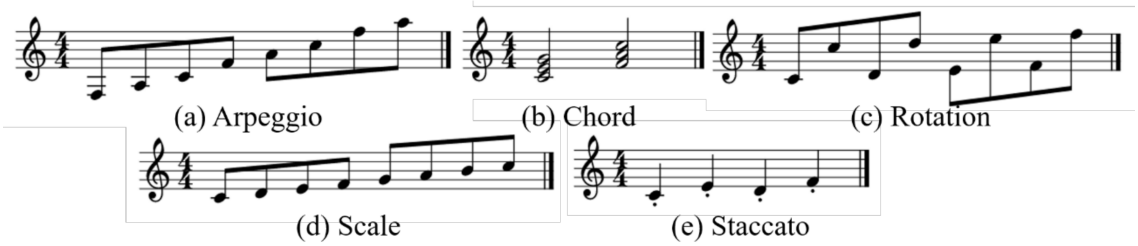


Figure 4. Representative piano playing techniques. (**a**) Arpeggio is a technique of playing the chords in rising or descending order from a low note to high; (**b**) chord is a set of more than one note; (**c**) rotation is a technique of playing two notes with a leap using the rotation of the front arm; (**d**) scale is a technique of playing the notes in one octave in order; (**e**) staccato is a technique of playing the notes with a short duration.

We also collected data of professional pianists having outstanding careers, such as international piano competition winners, and data of piano students playing for the hobby from the Youtube platform. A total of 360 videos were collected, and a total of 11,850 sequences were extracted by using OpenPose. The number of players in the piano playing videos was 13 (eight professionals and five amateurs), and the number of music pieces used was 17. It is important to note that the automatic annotation was performed according to the note composition that constitutes the technique. Here, note composition refers to the pattern of notes forming each technique. For example, if the interval between notes is 1 degree, it means the scale technique, and if multiple notes are stacked, it means the chord technique. The details of C3Pap is given in Table 1.

Table 1. Characteristics of Classic Piano Performance Postures of Amateur and Professionals (C3Pap) dataset.

Characteristic	Measure
Number of amateur pianists	8
Number of professional pianists	5
Number of music piece	17
Total sequence	11,850
Total length	1506 s
Mean clip length	4.18 s
Min clip length	1 s
Max clip length	14 s
Frame rate	29.97 fps
Audio	Yes

The C3Pap dataset has the following advantages over existing datasets proposed in [5,11–13], as shown in Figure 5. First, the proposed dataset has the largest total number of music pieces played by pianists, including music pieces from real performances. Second, the C3Pap dataset has the highest number of pianists compared with existing datasets, and the ratio of professional and amateur pianists is even. Moreover, the videos of the professional pianists in C3Pap include only the world's renowned pianists who have won numerous international piano competitions or contracted with major record labels. Third, the proposed dataset contains actual performance videos in diverse environments, unlike existing datasets that do not consider various situations of the pianist.

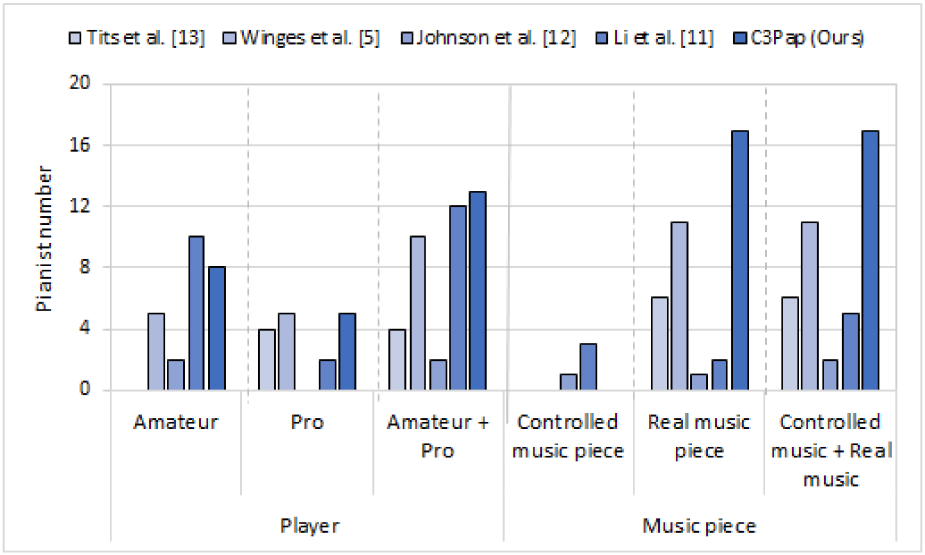


Figure 5. Advantage of the C3Pap dataset over existing datasets [5,11–13].

3.2. Feature Extraction

The feature extraction step for visual information proceeds as follows. The visual information of the skeleton is extracted using OpenPose, which enables us to extract skeleton information of the human body. According to previous research on piano playing, this study focused on the joints that should be viewed while playing piano rather than extracting all joints. Banowetz noted that the angle of the head and neck is related to the injury posture [20]. In [21], the angle of the head and neck is measured to analyze the level of improvement in piano performance. The authors of [5,22–24] measured the coordinates, angles, and silhouettes of the shoulders, spine, elbows, wrists, and fingers. The experimental results of [25] demonstrated that the pinky movements of amateur and professional pianists differed. Hence, this study used videos of pianists playing piano filmed from the right side

to observe the joints that are important in playing the piano. There is a total of 10 skeleton types to extract, including the head, neck, right shoulder, right elbow, right wrist, and four joints of the right pinky. Using OpenPose, the two-dimensional position information of each joint, and the attribute values of each coordinate can be extracted [18].

The feature extraction step for audio information proceeds as follows. After extracting the audio from the videos, we used the MFCC technique to extract the features of the audio information. MFCC is a state-of-the-art technique for extracting features from speech and is utilized in a variety of speech processing fields, such as instrument classification and speech classification. In the context of the proposed method, MFCC divides the audio information into certain sections and analyzes the spectrum for each section to extract the features. Once features for audio information are extracted, these are normalized to a value between −1 and 1.

3.3. Data Normalization

The data of the skeleton extracted from the visual information were normalized in two steps. In the first step, we perform a Procrustes transformation. As the C3Pap consists of piano playing videos collected from YouTube, the distance between the camera and the player and the direction of the camera differ in each video. If the data are not normalized, rather than training to classify postures according to the actual posture, the model will incorrectly train to classify postures according to the camera's location. To compensate for this, the skeleton coordinates of the entire frame are changed to match one frame. As such, by normalizing one frame to a standard, the camera direction and distance between the player and the camera can be unified to one standard.

Recall from Section 2 that skeleton extraction may lead to the following problems. The first problem is the distortion caused by the hand covering the piano, and the second problem is low video quality in the videos collected from YouTube. Both lead to issues with the skeleton recognition rate, ultimately causing the incorrect skeleton to be recognized or the skeleton to not be recognized at all, resulting in a null value. We consider this value as an outlier. Thus, in the second step, when the skeleton recognition rate decreases, the interquartile range (IQR) is used to compensate by removing outliers and replacing them with an average of adjacent values [26]. Specifically, we replaced the null values with the average of the frames before and after, in which the null value occurs using IQR. This procedure is repeated while increasing the left index and right index by increments of 1 until they are within the IQR.

3.4. Audio-Visual Tensor

To learn the relevance of sound and posture, it is important to preserve the identity of each data even if the data are fused. Unlike the previous method of digitizing data into a two-dimensional or three-dimensional data structure, this paper proposes mixing colors to identify each piece of data better. Specifically, we propose a novel multimodal data representation method that can express the relationships among data using an approach inspired by color characteristics. A single color can be expressed in terms of hue and brightness. For example, mixing red and gray results in gray-red. Because the colors are mixed by rules, even after mixing, it is possible to estimate what colors were mixed. If multimodal data are expressed using color and brightness, each feature can be preserved after fusion, and the relationship can be expressed by utilizing these features. As a result, an effective data representation method will help deep learning models perform decoding to a meaningful degree.

We represent visual information in grayscale and audio on the color scale. In the piano field, the lower and higher tones are different, and there is also a difference in the quality of sound produced by professional and amateur pianists. Thus, this complex set of information is represented in a 3D color scale. In contrast, we represent visual information in grayscale because we capture only the human skeleton's spatial information, which is relatively simple than audio information. For example, a professional pianist plays a chord with the wrist down to produce a rich sound. Thus, the audio tensor expressing the characteristics of the sound stores rich sound information in color scale (say red

color), and the visual tensor expressing the characteristics of the posture is a light grayscale expressing low wrist movement. The audio tensor is expressed in color by adjusting the CMY value, excluding K, and in the case of a visual tensor, it is expressed in grayscale by adjusting the K value, excluding CMY. Combining the two creates a red tensor with low brightness, which is classified as pro_chord. Thus, when using the proposed method, we can preserve the relationship between audio and visual information and enable the inter-modality features to be efficiently modeled.

Figure 6 shows the process of creating an audio-visual tensor. Here, we extract ten joints of the skeleton, where each joint has an x and y coordinate. Thus, the length of the horizontal axis of the visual tensor is 20 pixels. As the vertical axis was extracted at 30 FPS, the visual tensor is set to 30 pixels. To convert each coordinate value to grayscale, CMYK, a color representation method, is used. Grayscale can be produced by assigning the integer value to K. Accordingly, obtained coordinate values are multiplied by 100 and then converted to an integer. On the other hand, MFCC features are extracted from audio mp3 files obtained from the video. In this process, 1 s of audio is extracted to match the visual tensor and time sequence. Furthermore, an audio matrix is created based on the extracted MFCC features. The size of the audio matrix is set equal to that of the video matrix. If the value in the audio matrix is positive, then blue is assigned, and, if negative, then red is assigned. Finally, the CMY of the audio tensor (note without K) is combined with the K of the visual tensor to create an audio-visual tensor image.

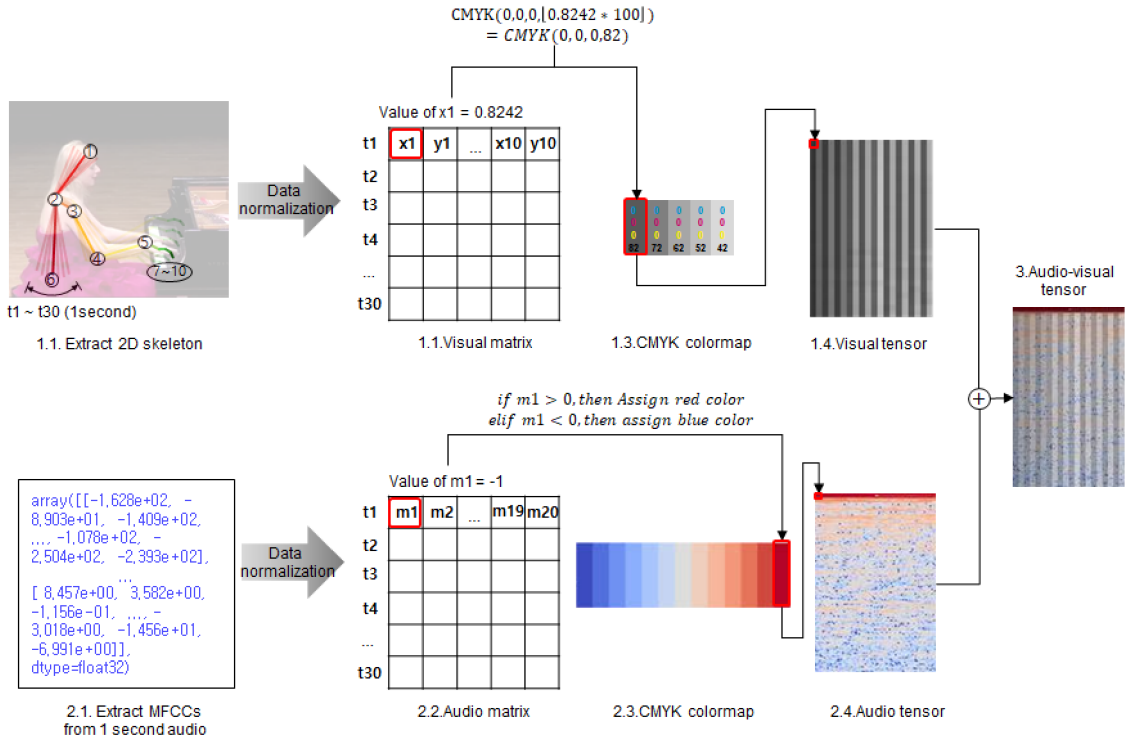


Figure 6. Construction process of audio-visual tensor in AV-TFN.

3.5. Model Training

Because the audio and video are represented as images in this study, the piano playing postures are classified using the Convolutional Neural Network (CNN), suitable for image analysis. The model was thus constructed as follows. Experiments using numerous hyperparameters demonstrated that kernel size and stride size influenced the performance of piano playing posture classification (refer to Section 4.2). Thus, the kernel size was set differently according to the input resolution. Specifically, when the resolution is 20 × 30, the kernel size is set to (1,1). When the resolution is 40 × 60, the kernel size is set to (2,2). When the resolution is 80 × 120 or 120 × 180, the kernel size is set to (6,6). The stride

size was also set differently according to the input resolution. When the resolution is 20 × 30 or 40 × 60, the stride size is set to 1. When the resolution is 80 × 120 or 120 × 180, the stride size is set to 6. Other conditions, including the number of layers, number of filters, batch size, dropout, and number of fully connected layers, did not impact posture classification accuracy. Accordingly, the number of filters was set to 32, the batch size was set to 200, and the epoch was set to 1000. For the optimizer, Stochastic Gradient Descent (SGD) with a learning rate of 0.001, momentum of 0.9, and decay of 1e-6 was used. Dropout was not used, and the number of fully connected layers was set to one. One CNN layer is used, and, after flattening the data to one dimension, ten dense unit layers are formed. Posture classification is then conducted using the Softmax function.

4. Performance Evaluation

In this section, we present performance evaluation. We first describe the implementation details and hyperparameter setting, and then we discuss the results of experiments.

4.1. Implementation Details

Table 2 shows a summary of the experiments conducted in this paper. To the best of our knowledge, there are no methods that use video and audio information to classify postures in the music domain. Thus, we have adopted several methods from different domains. Specifically, we compared the proposed AV-TFN with VN [7] and AN that utilize only visual and audio networks, respectively. Besides, we compared the proposed method with AVN_Concat [8] and AVN_Atten [9] that use audio-visual information with various data representation techniques.

Table 2. Summary of experiments and hyperparameters.

Ex		Parameters	
Ex. 1	Comparing performance of AV-TFN with other methods	Image resolution	80 × 120
		Kernel size	(6,6)
		Models	VN, AN, AVN-Concate, AVN-Atten, AV-TFN
		Tensor fusion ratio	50:50
Ex. 2	Estimating performance of model as image resolution increases	Image resolution	20 × 30, 40 × 60, 80 × 120, 120 × 180
		Epoch	1500
		Kernel size	(1,1), (2,2), (6,6)
		Models	AN, AVN-Concate, AV-TFN
		Tensor fusion ratio	50:50
Ex. 3	Estimating effect of tensor fusion ratio as tensor fusion ratio is varied	Image resolution	20 × 30
		Epoch	1500
		Kernel size	(1,1)
		Models	AV-TFN
		Tensor fusion ratio	0:100, 50:50, 100:0

Figure 7 demonstrates the structure of competing models. Each model was formed with the structure outputting the best performance (i.e., number of layers, number of filters, and optimizer). For VN and AN, the skeleton extracted from the video and MFCC extracted from the audio were used as input values. Through hyperparameter setting experiments (refer to Section 4.2), LSTM was selected to process time-series data for VN. Here, four LSTM layers and one hundred LSTM cells were used, where the epoch was set to 1000. As for AN, four CNN layers were used, where the number of filters was set to 32, the batch size was set to 200, the kernel size was set to (3,3), and the epoch was set to 100. Furthermore, after flattening the data to one dimension, ten dense unit layers are formed. In the last layer, the Softmax operation is performed to classify the posture. For both VN and AN, we use Adagrad optimizers with a learning rate of 0.01 and decay of 0.0.

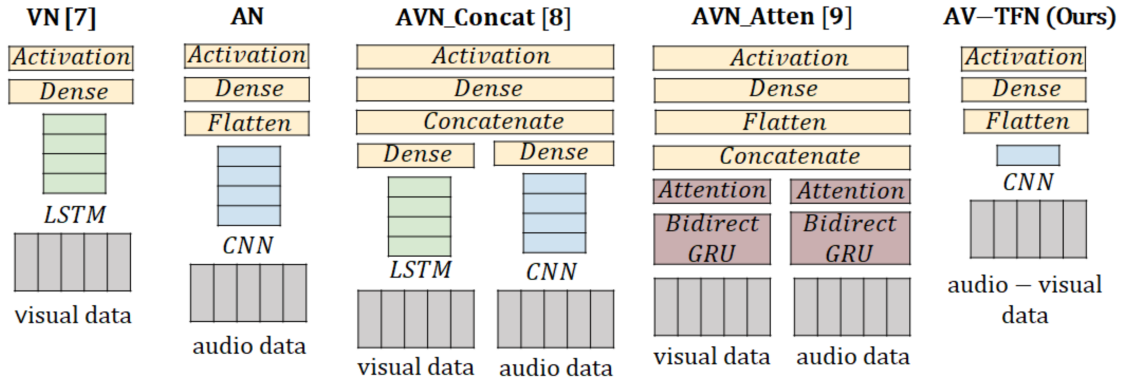


Figure 7. Structure of competing models.

We constructed AVN-Concat as follows. The input values for AVN-Concat are the skeleton data and MFCC. For this, LSTM and CNN were used, which are suitable for time series data processing and audio classification, respectively. For the visual information, four CNN layers were used for processing. Here, the number of filters was set to 32, the batch size was set to 200, the kernel size was set to (3,3), and the epoch was set to 100. Furthermore, after flattening the data to one dimension, the Softmax function is executed. For the audio information, four LSTM layers comprising 100 cells are used for processing. Furthermore, after flattening the data to one dimension, the Softmax function is executed. Each stream of CNN and LSTM is concatenated, and ten dense unit layers are formed. In the last layer, Softmax operation is performed one more time to classify the postures. We use SGD optimizers for AVN-Concat, where a learning rate was set to 0.001, decay was set to 1e-6, and momentum was set to 0.9. As for AVN-Atten, the input data are concatenated into one dimension through a bi-direction Gated Recurrent Unit (GRU) layer with 200 attention units. Furthermore, AVN-Atten performs classification of postures using a dense layer and Adam optimizers, where a learning rate was set to 0.001, the beta_1 was wet to 0.9, the beta_2 was wet to 0.999, and the decay was set to 0.0.

Regarding the training and test ratio, we used 80% of the total data as a training dataset and 20% as test data, where data were randomly chosen. In experiments 1–3 (Table 2), the same training dataset and test data were used.

The experimental environment is as follows. For the CPU, an Intel Xeon CPU E5-2620 v3 2.40GHz was used, and, for the graphics cards, Titan XP and GeForce GTX TITAN were used. Keras and TensorFlow were used as the development environment. Ubuntu 16.04 was used as the operating system, and Python3 was selected as the development language.

4.2. Hyperparameter Setting

Figure 8a shows a graph of the F1 score measurements with various posture classification algorithms. In Figure 8a, x and y axes indicate the algorithm types and F1 score, respectively. The algorithm types used in the experiments are GRU, LSTM, and 2D Convolutional LSTM (2DConvLSTM), which are suitable for posture classification. The number of layers and cells for both GRU and LSTM is set to 3 and 100, respectively. From the graph, we can observe that LSTM exhibited the highest F1 score among the three posture classification algorithms. In general, LSTM has higher classification accuracy in longer sequences than GRU. Considering that 30 frame rate of long data sequences were used in our experiments, LSTM output higher accuracy than GRU. Moreover, considering that we use one-dimensional data, LSTM demonstrates a better performance than 2DConvLSTM, which is more suitable for two-dimensional data. Therefore, LSTM was selected as a method to perform VN.

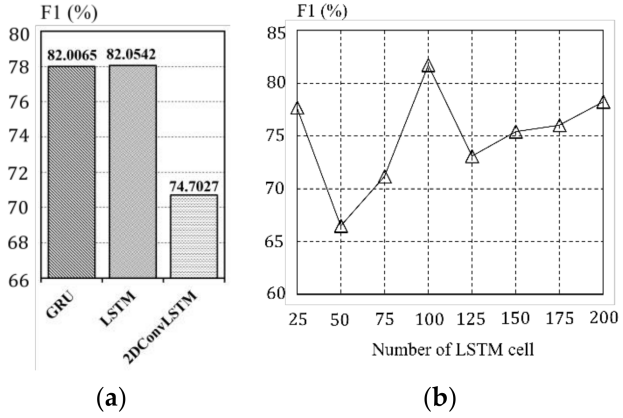


Figure 8. Experiment results with various hyperparameter setting: (**a**) VN algorithms; (**b**) number of LSTM cell.

Figure 8b shows the F1 score changes for posture classification with increasing LSTM cells. In Figure 8b, x and y axes indicate the number of LSTM cells at an increment rate of 25 and the F1 score, respectively. From the graph, we can observe that the F1 score was the highest with 100 cells when the number of LSTM layer was set 1. The line showed an upward trend as the number of cells increased; however, a regular trend was not observed, and the F1 score decreased again with 100 or more cells. Accordingly, the number of LSTM cells in the LSTM-based posture classification model was set to 100.

Figure 9 demonstrates the effect of kernel and stride sizes on AV-TFN performance. Specifically, Figure 9a shows the changes in the F1 score for posture classification with increasing CNN kernel size. In Figure 9a, x and y axes indicate the kernel size and the F1 score, respectively. We set the input resolution to 80 × 120 and the stride size to 5. From the graph, we can observe that as the kernel size increased, the F1 score also gradually increased, exhibiting the highest score at (6, 6) and (7, 7). On the other hand, Figure 9b shows the changes in the F1 score for posture classification with increasing CNN kernel size, which is not square. When we set input resolution and stride size identical to the previous experiment, we can observe that the F1 gradually increased, exhibiting the highest score (4, 5), (5, 4). Furthermore, Figure 9c shows the F1 score changes for posture classification as the stride size increases. In Figure 9c, x and y axes indicate the stride size and the F1 score, respectively. We set the input resolution to 80 × 120 and the kernel size to (6, 6). The experiment results indicate that when the stride size was 1–4, the F1 score was less than 10%, and when the stride size was set to 5 or more, the F1 score increased to at least 80%.

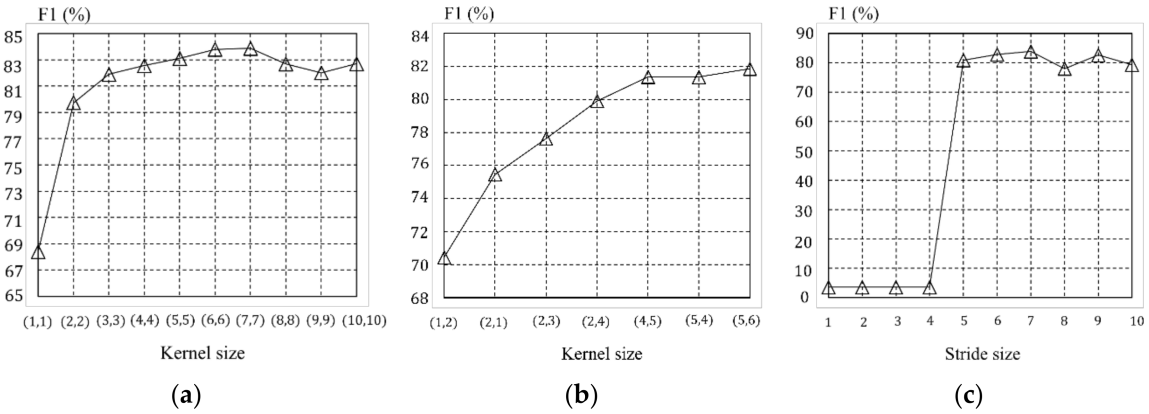


Figure 9. Experiment results with various hyperparameter setting: (**a**) and (**b**) kernel size; (**c**) stride size.

4.3. Experiments and Results

Experiment 1. *Comparing performance of AV-TFN with other methods.*

Table 3 presents the experimental results, demonstrating that the proposed method achieves the highest accuracy, precision, recall, F1 score, and lowest loss. From the table, we can also observe that the proposed method achieves the fastest training time and fastest test time. AV-TFN outperforms VN as it uses audio information to compensate for cases in which the posture classification accuracy may be degraded because of deteriorating video quality and rapid hand movement. On the other hand, AV-TFN outperforms AVN-Concat and AVN-Atten owing to a more efficient audio-visual fusion strategy. That is, the existing multi-modal methods simply concatenate audio-visual features, which does not allow the intra-modality features to be modeled efficiently. Recall from Section 3 that, in AV-TFN, the audio tensor is expressed on the color scale by adjusting the CMY value, excluding K, whereas the visual tensor is expressed in grayscale by adjusting the K value, excluding CMY. Combining the two creates a tensor that preserves the relationship between the audio and visual information.

Table 3. Experiment results on comparison of AV-TFN WITH EXISTING METHODS.

	Accuracy	Precision	Recall	F1 Score	Error Rate	Training Time	Test Time
VN [7]	84.9829	83.6057	81.7531	82.0542	0.6118	1.0963	0.6041
AN	80.137	82.8966	80.6059	81.0313	0.9951	1.7541	3.4616
AVN-Concate [8]	76.0274	79.8384	72.9123	73.53	0.9951	5.1612	5.2532
AVN-Atten [9]	81.57	84.8317	75.3428	75.0957	1.0746	5.1982	4.0455
AV-TFN (Ours)	87.7133	86.1329	84.7091	84.9085	0.3943	0.3454	0.2799

Experiment 2. *Estimating performance of model as image resolution increases.*

We have also performed more detailed experiments to measure the effect of image resolution on classification accuracy and to find the optimal resolution (refer to Figure 10). This experiment compares the F1 scores, training time, and test time of the AN, AVN, and AV-TFN as produced tensor resolution increases. Here, a resolution is determined by regulating the MFCC feature (sampling rates and MFCC coefficients). From the graphs in Figure 10a,b, we can observe that AV-TFN showed the highest F1 score and lowest error rate at all resolutions. This is because AV-TFN effectively expresses the relationship of data by using color features and maintains the existing features even after fusion. More specifically, Figure 10a shows the changes in the F1 scores of the three models as the image resolution increases. From the graph, we can observe that the F1 score increases as the input image resolutions increase. We can also observe that all models tended to improve the F1 score until the image resolution was 80×120, and then fell. We can conclude that the image resolution of MFCC has a considerable influence on all models. Therefore, if the resolution is low, visual and audio information is not sufficiently included, which results in a reduced F1 score. On the other hand, when audio features are extracted with excessive detail, such as 120×180, noise increases, degrading classification accuracy.

Figure 10c,d show the changes in the training time and testing time of the three models as the input image resolution increases. From the graphs, we can observe that AN and AVN exhibited higher training time and test time as the input image resolution increased. In the case of AVN models, the network processing the auditory information and the network processing the visual information must run in sequence, thus lengthening the test and training time. For AV-TFN, the training time and test time barely increased even with increasing input image resolution. In AV-TFN, the two types of input values are represented through one data structure that is processed as one network, thus maintaining relatively low training and test time.

Table 4 demonstrates the results of the experiments as image resolution increases (i.e., in terms of superiority by percentage points). From the table, we can observe the AV-TFN exhibits an F1 score of 9.1161 percent points (on average) higher than AN, and 16.7775 percent points (on average) higher than AVN models.

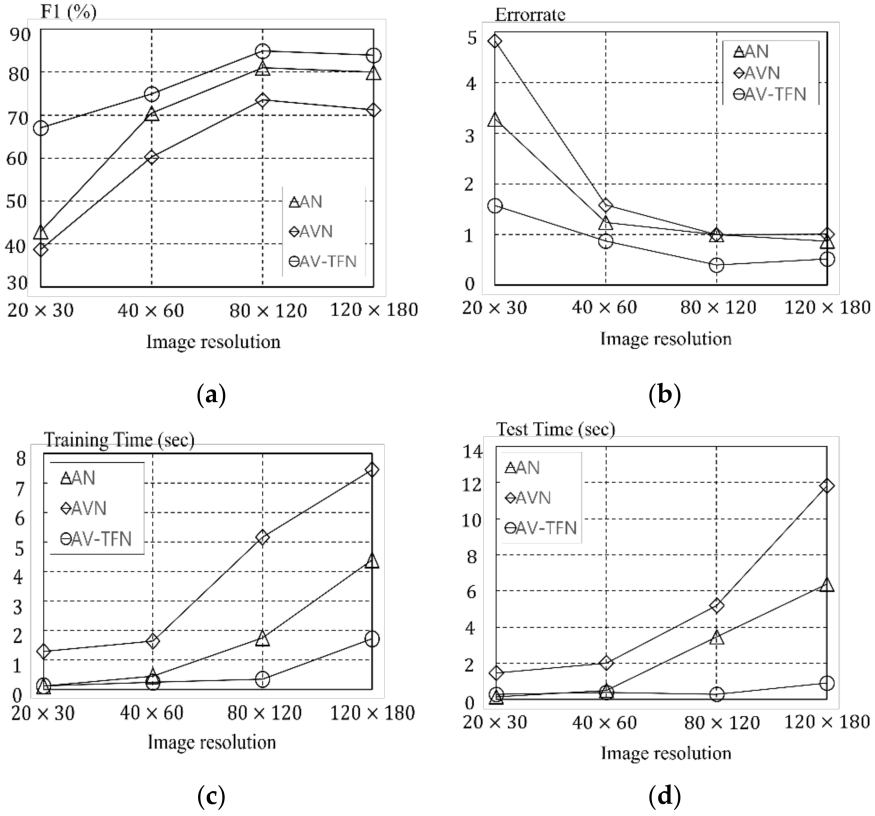


Figure 10. Experiment results of AN, Audio-Visual Network (AVN), AV-TFN as image resolution increases: (**a**) F1 score; (**b**) error rate; (**c**) training time; (**d**) test time.

Table 4. Superiority of AV-TFN compared with AN and AVN.

Image Resolution	AN vs. AV-TFN				AVN vs. AV-TFN			
	F1 Score	Error Rate	Training Time	Test Time	F1 Score	Error Rate	Training Time	Test Time
20 × 30	24.136	1.7118	0.008	0.0031	28.3329	3.251	1.9513	0.944
40 × 60	4.4681	0.3704	0.5471	0.7885	14.6703	0.7126	2.4798	1.4651
80 × 120	3.8772	0.6008	1.4087	3.1817	11.3785	0.6008	4.8158	4.9733
120 × 180	3.9825	0.3508	3.3398	5.668	12.7281	0.4891	9.2675	14.9646
Average	9.116	0.7447	1.3259	2.4103	16.7775	1.2634	4.6286	5.5868

Experiment 3. *An experiment on estimating effect of tensor fusion Ratio.*

An experiment was conducted to test the synergy of audio and video information. As we mentioned in Section 3, when creating an audio-visual tensor, we mixed the colors of the audio and visual tensors and set the ratio to 0:100, 50:50, and 100:0. Here, the Tensor Fusion Rate (TFR) 0:100 means the audio tensor color ratio is 100%, while the visual tensor color ratio is 0%. Table 5 demonstrates changes in the F1 score for posture classification according to the TFR of the visual and audio tensors. Experimental results show that the F1 score in the case of 50:50 improves by about 5.4052 percentage points and 13.0826 percentage points when the two are combined than in the case of 0:100 and 100:0. This indicates that audio-visual tensors are more synergistic than they are used individually.

Table 5. Performance of AV-TFN as Tensor Fusion Rate (TFR) varied.

IR	TFR	Accuracy	Precision	Recall	F1 Score
20 × 30	0:100	64.7917	64.2071	61.1152	61.636
	50:50	70.2083	69.4851	66.4125	67.0412
	100:0	62.7083	58.9199	55.3345	53.9586

5. Conclusions

In this paper, we have proposed an audio-visual tensor fusion network (simply, AV-TFN) for piano playing posture classification. Unlike existing studies that used only visual information, the proposed method uses audio information to improve the accuracy in classifying the postures of professional and amateur pianists. We have compared the proposed method with its variants: VN (Visual Network), AN (Audio Network), and AVN (Audio-Visual Network). The experiment results demonstrate that AV-TFN outperforms existing studies and, thus, can be effectively used in the classification of piano playing postures. This study has the following limitations. First, 3D coordinate values were not used as input when classifying the piano playing posture, and only 2D coordinate values were used. Second, only videos of pianists playing piano filmed from the right side were used. Future studies will compare the performance of posture classification using 2D and 3D coordinates. Furthermore, video data filmed not only from the right side but also from the left, above, and behind will also be tested.

Author Contributions: S.-H.P. designed the algorithm and developed the proposed algorithm and writing of the paper. Y.-H.P. shared his expertise with regard to the overall review of this paper and supervised the entire process. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology [NRF-2018R1D1A1B07046550]. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology [NRF-2019R1A6A3A13096032]. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.2016-0-00406, SIAT CCTV Cloud Platform).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bragge, P.; Bialocerkowski, A.; McMeeken, J. A systematic review of prevalence and risk factors associated with playing-related musculoskeletal disorders in pianists. *Occup. Med.* **2006**, *56*, 28–38. [CrossRef] [PubMed]
2. Neninger, C.R.; Sun, Y.; Lee, S.H.; Chodil, J. A complete motion and music capture system to study hand injuries among musicians. In Proceedings of the International Conference on Emergency Management & Robotics for Hazardous Environments, Knoxville, TN, USA, 8–10 August 2011; pp. 1–11.
3. Zandt-Escobar, A.V.; Caramjaux, B.; Tanaka, A. PiaF: A tool for augmented piano performance using gesture variation following. In Proceedings of the International Conference on New Interfaces for Musical Expression, London, UK, 30 June–4 July 2014; pp. 167–170.
4. Hadjakos, A. Pianist motion capture with the Kinect depth camera. In Proceedings of the 9th Sound and Music Computing Conference, Copenhage, Denmark, 11–14 July 2012; pp. 303–310.
5. Winges, S.A.; Furuya, S. Distinct digit kinematics by professional and amateur pianists. *Neuroscience* **2015**, *284*, 643–652. [CrossRef] [PubMed]
6. Park, S.H.; Nasridinov, A.; Park, Y.H. A kinect-based piano education system for correction of pianist posture. *Asia Life Sci.* **2015**, *12*, 571–586.
7. Park, S.H.; Ihm, S.Y.; Nasridinov, A.; Park, Y.H. A Feasibility Test on Preventing PRMDs Based on Deep Learning. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 10005–10006.
8. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL-based multimodal emotion recognition and sentiment analysis. In Proceedings of the IEEE 16th International Conference on Data Mining, Barcelona, Spain, 12–15 December 2016; pp. 439–448.
9. Akhtar, M.S.; Chauhan, D.S.; Ghosal, D.; Poria, S.; Ekbal, A.; Bhattacharyya, P. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 370–379.
10. Mora, J.; Lee, W.S.; Comeau, G. 3D visual feedback in learning of piano posture. In Proceedings of the International Conference on Technologies for E-Learning and Digital Entertainment, Hong Kong, China, 11–13 June 2007; pp. 763–771.

11. Li, M.; Savvidou, P.; Willis, B.; Skubic, M. Using the kinect to detect potentially harmful hand postures in pianists. In Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; pp. 762–765.
12. Johnson, D.; Dufour, I.; Damian, D.; Tzanetakis, G. Detecting pianist hand posture mistakes for virtual piano tutoring. In Proceedings of the 42nd International Computer Music Conference, Utrecht, The Netherlands, 12–16 September 2016; pp. 168–171.
13. Tits, M.; Tilmanne, J.; D'Alessandro, N.; Wanderley, M.M. Feature extraction and expertise analysis of pianists' Motion-Captured Finger Gestures. In Proceedings of the 41st International Computer Music Conference, Denton, TX, USA, 25 September–1 October 2015; pp. 1–4.
14. Shlizerman, E.; Dery, L.; Schoen, H.; Kemelmacher-Shlizerman, I. Audio to body dynamics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7574–7583.
15. Ueno, K.; Frukawa, K.; Nagano, M.; Asami, T.; Yoshida, R.; Yoshida, F.; Saito, I. Good Posture Improves Cello Performance. In Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Hong Kong, China, 29 October–1 November 1998; pp. 2386–2389.
16. Morency, L.; Mihalcea, R.; Doshi, P. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. In Proceedings of the 13th International Conference on Multimodal Interfaces, Suzhou, China, 14–18 October 2019; pp. 169–176.
17. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 1103–1114.
18. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 7291–7299.
19. Sándor, G. *On Piano Playing: Motion, Sound and Expression*; Schirmer Books: New York, NY, USA, 1981; pp. 37–140.
20. Banowetz, J. Piano-Related Musculoskeletal Disorders: Posture and Pain. Doctoral Dissertation, University of North Texas, Denton, TX, USA, May 2013.
21. Beacon, J.F.; Comeau, G.; Payeur, P.; Russell, D. Assessing the suitability of Kinect for measuring the impact of a week-long Feldenkrais method workshop on pianists' posture and movement. *J. Music Technol. Educ.* **2017**, *10*, 51–72.
22. Payeur, P.; Nascimento, G.M.G.; Beacon, J.; Comeau, G.; Cretu, A.M.; D'Aoust, V.; Charpentier, M.A. Human gesture quantification: An evaluation tool for somatic training and piano performance. In Proceedings of the IEEE International Symposium on Haptic, Audio and Visual Environments and Games, Houston, TX, USA, 23–26 February 2014; pp. 100–105.
23. Willis, B.; Li, M.; Skubic, M. Assessing injury risk in pianists: Using objective measures to promote self-awareness. *MTNA e-J.* **2017**, *9*, 3–17.
24. Hadjakos, A.; Aitenbichler, E.; Mühlhäuser, M. The Elbow Piano: Sonification of Piano Playing Movements. In Proceedings of the 8th International Conference on New Interfaces for Musical Expression, Genova, Italy, 5–7 June 2008; pp. 285–288.
25. Furuya, S.; Flanders, M.; Soechting, J.F. Hand kinematics of piano playing. *J. Neurophysiol.* **2011**, *106*, 2849–2864. [CrossRef] [PubMed]
26. Sadeghzadehyazdi, N.; Batabyal, T.; Dhar, N.K.; Familoni, B.O.; Iftekharuddin, K.M.; Acton, S.T. GlidarCo: Gait recognition by 3D skeleton estimation and biometric feature correction of flash lidar data. *arXiv* **2019**, arXiv:1905.07058.