EXPLORING CONVOLUTIONAL NEURAL NETWORK MODELS FOR MULTIMODAL CLASSIFICATION OF EXPRESSIVE PIANO PERFORMANCE

Anna-Maria CHRISTODOULOU[1,3], **Sagar DUTTA**[1,3], **Olivier LARTILLOT**[1,3], **Kyrre GLETTE**[2,3], and **Alexander Refsum JENSENIUS**[1,3]

[1]*Department of Musicology*, **University of Oslo**, Oslo, Norway
[2]*Department of Informatics*, **University of Oslo**, Oslo, Norway
[3]*RITMO, Center of Excellence for Rhythm, Time, and Motion*, **University of Oslo**, Oslo, Norway

ABSTRACT

This paper addresses improving performance analysis by automating the recognition of expressive performance styles. We propose a multimodal fusion approach integrating audio, video, and motion data. We demonstrate the effectiveness of our approach by utilizing convolutional neural network (CNN) models. Training is done on a classical piano dataset of 211 excerpts containing audio, video, MIDI, and motion capture data. The results highlight the robustness of the CNN models; they achieve high accuracy even when trained on a limited dataset. Our study contributes to advancing the field of performance analysis by applying deep learning techniques to multimodal data.

1. INTRODUCTION

Musicians use many types of body motion while performing, ranging from sound-producing actions to expressive gestures that may not directly affect the musical sound [1]. Here, we are interested in music performance styles, especially those related to piano performance. Music performances are multimodal in nature. We define multimodality as the combination of diverse data types that offer complementary information to the processing task, typically arising from perceived relationships between sound-producing actions and the resulting sounds. However, they may also be evoked from only auditory or visual modalities. Musical actions vary and can help shape the musical phrasing, express emotions, or communicate a structural change in the piece. Interestingly, even subtle actions can be perceived as significant and influence the conveyance of expressive elements; moving an eyebrow or gesturing with a finger can be powerful artistic expressions.

Humans can easily observe and understand the many layers of musical actions in performance, and professional musicians, educators, and music researchers have a highly developed sense of such bodily nuances. However, many challenges arise when training machine learning systems to identify musical actions in a dataset. Convolutional Neural Networks (CNNs), a type of deep learning model particularly effective in processing visual and spatiotemporal data, have shown great promise in this area. CNNs can automatically learn and extract features from input data, making them suitable for recognizing complex music performance patterns. Such an *automated expressive music performance recognition* task is central when analyzing music performances.

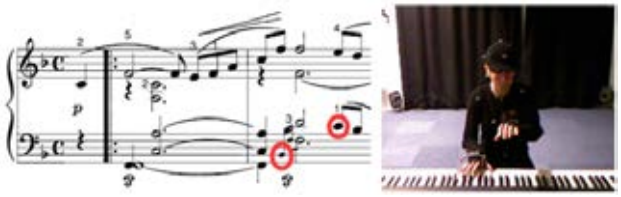

Figure 1. A left-hand phrase from Traumerei (left), between the circled notes, and a still image from video (right) showing normal tempo, exaggerated expression. [2]

Our research is driven by the goal of exploring the potential of integrating multimodal approaches and deep learning models in automated expressive music performance recognition during music performances (Fig. 2). We specifically address the challenges of combining audio, video, and motion information while identifying unexplored possibilities for advancing the recognition and analysis of expressive music performance. We aim to answer the following research questions:

- Can a 1D CNN accurately detect spatiotemporal information from piano performances?
- How can we work with multimodal CNNs for a music performance classification task, and how many dimensions are necessary?
- Is there an effect in accuracy between unimodal and multimodal approaches?

In the following sections, we analyze related work and discuss the effectiveness of multimodal approaches in music recognition tasks. We provide details about the dataset used, as well as the architecture of our model, and share our observations on the obtained results. Through this initial exploration, we seek to establish the groundwork for future strides in multimodal music performance recognition. [1]

[1] Our pre-trained models and our code are publically available with an

2. RELATED WORK

Music experiences are inherently multimodal. In a concert, we *see* musicians playing on their instruments, *hear* the resultant musical sound, *smell* people around us and the environment, *feel* the vestibular motion when moving or dancing to music or the sound waves reaching our bodies in a loud concert [3]. Music may even change the *taste* of the food we are eating [4]. Similarly, when performing, we *touch* the instrument or *move* our body to the music, we *hear* the sound and *see* the sheet, the instrument, or the conductor instructing us how to play [5]. Thus, to understand musical experiences, it is necessary to employ analysis techniques that allow for studying the complexity at hand in a complex and multimodal manner. However, significant gaps remain, particularly in integrating multiple modalities to capture and characterize musical expressive performances comprehensively.

Past research has shed light on how individuals perceive and represent musical phrases through bodily movements [6]. Understanding humans' mapping strategies provides insights into how we interpret and express music through body movements. Such findings have implications in various domains, such as music performance, education, and human-computer interaction, adding to our understanding of individual and cultural variations in gestural interpretations of music. This can also contribute to our understanding of the embodiment of music and show the importance and relevance of research in music expression analysis, as well as related fields.

Sarasua showed that multimodal music analysis outperforms unimodal approaches in various music processing tasks, including music gesture recognition [7]. They explored context-aware gesture recognition of a conductor's gestures, incorporating audio and motion capture data. This involved clustering similar gestures based on extracted motion capture and audio features. The study showcased the potential of analyzing and interpreting the conductor's expressivity during musical performance. However, video data, which could provide vital visual cues, was not included.

Wu et al. [8] proposed a classifier combination framework for action recognition using audio and motion capture. Their approach involved extracting audio features, such as Mel-frequency cepstral coefficients (MFCCs), and motion features related to joint angles and accelerations. By combining these features, they achieved an accuracy improvement of 6% compared to previous single-modal approaches.

Chang et al. [9] focused on modernizing music recording using classification techniques. Their automated analysis pipeline for piano playing sessions utilized multiple data types, including text, audio, video, and electroencephalography (EEG). Their approach involved extracting features such as pitch salience, dynamics, and spectral flux from the audio signals and capturing video data of hand motion.

Outside music-related research, Nandakumar et al. [10] presented a novel multimodal action recognition approach based on skeletal joint positions. Their approach included a spotting module to detect gesture instances and a recognition module to classify the gesture types. They achieved an average recognition accuracy of 87%.

In summary, while previous studies have significantly contributed to action recognition, gaps remain, particularly in integrating audio, video, and motion data to characterize expressive music performances. By incorporating multiple modalities in the dataset and employing advanced machine-learning techniques, we aim to open new musical interpretations and gain a deeper understanding of the relationships between music sound and motion.

3. METHODS

This paper used a multimodal music dataset to study expressive piano performances. The dataset included audio, video, and motion capture. After a thorough data preprocessing, we implemented three classification rounds, one for each data type. We then concatenated the separate representations into a Multimodal CNN classifier. Throughout our experiment, we also integrated regularization techniques to improve generalization and convergence.

3.1 Dataset Description

For this paper, we used a multimodal piano dataset from Sarasua et al. [2] to analyze expressive music performance. The dataset is available under the CC BY-NC-SA license [11] and comprises audio, video, motion capture, and MIDI data. The participating pianists performed an excerpt of Robert Schumann's Träumerei (Kinderszenen Op.15 No.7) (Fig. 3).

The dataset is of two pianists (*pianist01* and *pianist02*) who perform the excerpt in three conditions (normal, slow, and fast), with and without a metronome. Each condition was recorded three times. One exception was that pianist01 had two recordings for the conditions metronome, exaggerated performance, and normal speed. Thus, the complete dataset contains 211 samples for the two pianists. In each of these conditions, their expressive performance can be categorized under the following five expressive styles:

- **Normal:** a standard performance without specific emphasis or modifications.
- **Still:** the pianists aimed to minimize their actions, keeping their bodies and hands as still as possible while playing, while the resulting sound was consistent with minimal variation.
- **Exaggerated:** exaggerated actions and expressions, where the pianists intentionally emphasize certain changes in dynamics and articulation.
- **Finger Legato:** smooth hand actions and connected sounds by emphasizing finger legato technique without noticeable breaks or pauses.
- **Finger Staccato:** short and distinct sounds using a finger staccato technique and sharp, quick hand actions.

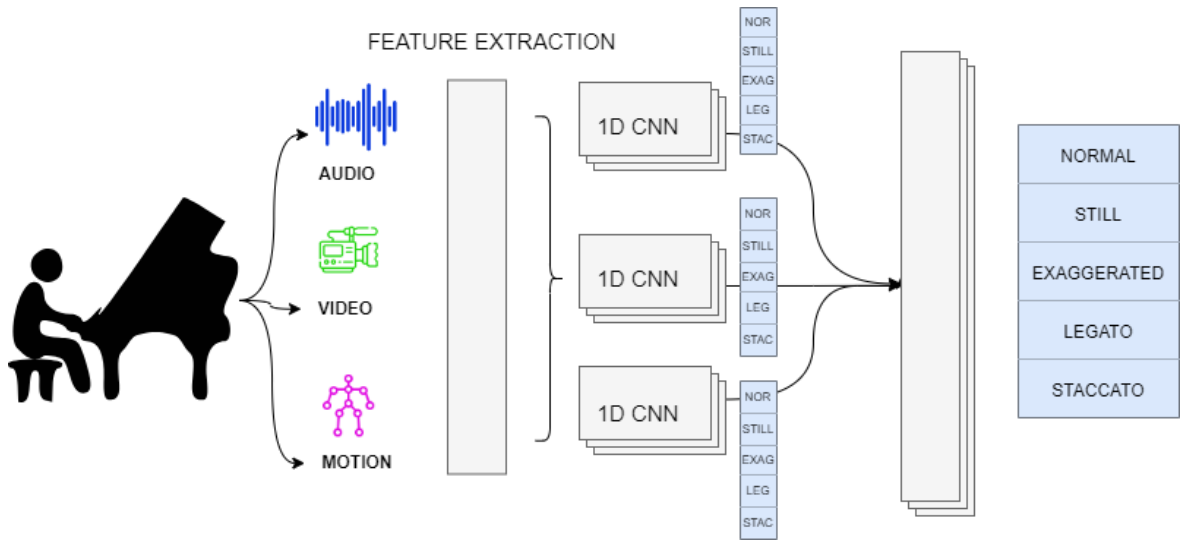

Figure 2. An overview of the Multimodal Fusion Method. The model receives three types of input data from the pianist's recordings (features extracted from audio, video, and motion capture). These inputs are processed and fed into a 1-dimensional CNN, each providing a classification result. The representation layers of the individual CNNs are then fed into the Multimodal CNN for the final classification output.

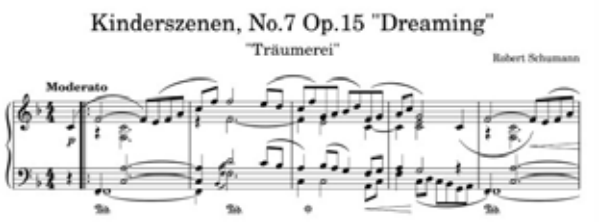

Figure 3. The excerpt from Kinderszenen Op.15 No.7 from Robert Schumann's Träumerei used in the dataset recordings we analyzed.

The audio was recorded at a sampling rate of 44.1 kHz, the video at 30 frames per second (fps), and motion capture data at 100 Hz. The motion capture data includes 22 features (position of body parts), but only two are retained for this project: the right and left hand in three dimensions. The recorded performances range from 8 to 48 seconds, the latter being the slow-speed recording. This provided a diverse data set for analysis across modalities.

3.2 Feature Extraction

Features were extracted from audio (RMS, spectral contrast), video (motiongrams), and motion capture data (velocity, acceleration). All of the features were padded to the maximum length of the input data (different for each data type). Padding is essential in data processing for machine learning models, particularly when dealing with sequences of different lengths. Consistent input shapes are maintained by padding data to a maximum length. This enables seamless integration with models that expect fixed-size *tensors*, algebraic objects that describe a multilinear relationship between sets of algebraic objects related to a vector space. The features were also concatenated with the corresponding labels attached to each sample and implemented one-hot label encoding. This means the labels for our five different expressive categories were transformed from numerical values between 0 and 4 to a sequence of 0 seconds. By implementing one-hot label encoding and concatenating the labels to each sample, we ensure that the labels are adequately represented for model compatibility, loss calculations, evaluation metrics, and ease of integration. This enables effective training and evaluation of machine learning models in our classification task.

3.2.1 Audio

Two features were extracted from the audio, using Librosa [12] in Python: the amplitude level and the spectral contrast [13]. The average amplitude level, calculated as the root mean square (RMS), reflects the audio signal's overall energy content and indicates the dynamic nature of the performance. Variations in touch, articulation, and dynamics manifest as distinct RMS values, allowing for differentiation between soft and gentle or forceful and intense actions. The spectral contrast captures disparities in spectral energy across frequency bands. This feature discerns the unique spectral profiles of diverse playing techniques, such as staccato and legato, marking the significance of frequency-specific variations in expressive piano playing.

3.2.2 Video

We extracted the vertical and horizontal motiongrams from the video data. Motiongrams represent the intensity of motion in a video. They are obtained by calculating the pixel differences between consecutive frames, both horizontally and vertically [3]. Let M be a motiongram array of shape $(w+1, n)$, where w represents the width of the motiongram and n is the total number of frames. We can compute the 1D feature vector F of length $w+1$ as follows:

$$F_j = \frac{1}{n}\sum_{i=1}^{n} M_{ji}, \quad j = 1, 2, \ldots, w+1$$

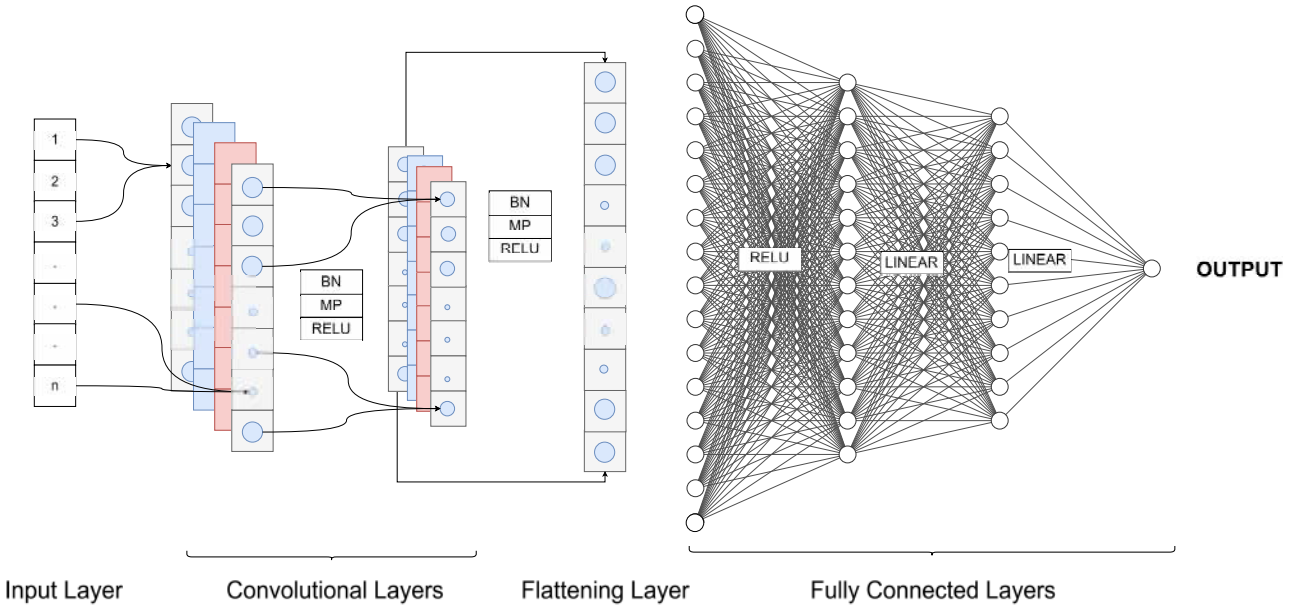

Figure 4. Illustration of the 1D Convolutional Neural Network (CNN) architecture designed for analyzing audio, video, and motion capture data. The architecture consists of three 1-dimensional convolutional layers (conv1, conv2, conv3) with 16, 32, and 64 output channels, respectively, followed by batch normalization and max pooling. The flattened output is then passed through fully connected layers (fc1, fc2, fc3, fc4, fc5) with the final softmax output layer supporting classification into different classes based on the unique characteristics of the input data.

The equation represents the vertical averaging process, where F_j is the j-th element of the 1D feature vector F and M_{ji} represents the motion value in the j-th column and i-th frame of the motiongram array M. These computed features F capture the overall motion patterns in the video and can be used for various analysis or modeling tasks.

We averaged the horizontal and vertical motiongrams in one dimension. This resulted in a concise visualization of the motion intensity throughout the video, helping us compress our motion information while dimensionalizing it and preserving the relevant temporal patterns.

3.2.3 Motion Capture

The motion capture data represents the instantaneous position (in 3 axes) of the participants' right and left hands. We extracted velocity and acceleration from these, using NumPy's gradient function to compute the derivatives of the positional data. Each hand's resulting velocity and acceleration arrays were then stacked to form a feature matrix, which was saved as a NumPy binary file for further analysis and processing. This enabled comprehensive motion data to be extracted from the original motion capture files. To check for overfitting, we visualized the resultant vectors of the motion for the left hand and right hand.

3.3 Model Description

CNNs demonstrated remarkable performance in capturing spatial and temporal patterns within data, making them suitable for analyzing time-series data [14]. A 1-dimensional Convolutional Neural Network (1D CNN) was built using Python for each modality (audio, video, and motion capture) (Fig. 4). The architecture comprised three 1-dimensional convolutional layers (conv1, conv2, conv3) with 16, 32, and 64 output channels, respectively, each followed by batch normalization [15] and max pooling [16]. The convolutional layers had a kernel size of 2, a stride of 1, and padding of 1 to maintain the input dimensions. The output of the convolutions was flattened and passed through fully connected layers (fc1, fc2, fc3, fc4, fc5) with 256, 128, 64, 32, and 5 (according to the number of classes) output nodes, respectively. Batch normalization was applied after each fully connected layer. ReLU activation functions were used after each layer to introduce non-linearity. The consecutive linear layers were designed to progressively transform and refine the feature representations, capturing complex patterns and relationships that might not have been possible with a single linear layer. Each layer added a level of abstraction, allowing the model to learn more nuanced features.

The 1D CNN allowed the exploitation of the temporal nature of audio data. Using convolutional layers, the model could learn and detect temporal relationships between different sound segments, extracting important features for classification. The hierarchical representations obtained by the convolutional layers captured both local audio features and global audio characteristics, enabling the model to effectively distinguish between different audio classes.

Similarly, it enabled the network to capture temporal dynamics in video analysis. The model discerned motion information contributing to different classes' visual representation by convolving along the temporal dimension. Combining convolutional layers and batch normalization aided in learning robust visual features, while the fully connected layers captured complex relationships between these features.

A 1D CNN was also well-suited for processing the sequential nature of motion capture sequences. The convolutional layers learned spatial dependencies over time, capturing joint position and motion changes throughout the sequence. This enabled the model to extract discriminative motion features for classification.

Batch normalization after each convolutional layer stabilized the learning process by ensuring consistent feature scaling and reducing the impact of covariate shifts. The rectified linear units (ReLU) introduced non-linearities, allowing the model to learn complex representations.

The final softmax output layer was designed to support classifying inputs into different classes, potentially offering predictions based on audio, video, and motion capture data. It was the last activation function of a neural network, normalizing the network's output to a probability distribution over predicted output classes. By utilizing 1D CNNs for each modality, the model aimed to exploit the unique characteristics and temporal dependencies inherent in audio, video, and motion capture data, potentially yielding improvements in classification performance.

3.4 Training process

In the two-stage training process, the data is utilized as follows: in the first stage, the CNN is trained on the individual modality data (audio, video, or motion capture) independently to learn modality-specific characteristics and temporal dependencies. In the second stage, the pre-trained CNNs for each modality are combined and further trained on the multi-modal data to capture cross-modal correlations and enhance classification performance. This two-stage approach leverages the unique temporal dynamics of each modality and effectively integrates information from multiple data sources.

In more detail, we began training by developing distinct CNN models for individual modalities, including audio, video, and mocap data. We had 211 vectors for each modality, with lengths of 431, 1225, and 9794, respectively. Each CNN model is trained to autonomously extract relevant features and embeddings from its respective input data. Upon training completion, the trained CNN models are utilized to generate embedding vectors from the input data of their corresponding modalities by extracting the outputs from the convolutional layers of each CNN. These outputs are the learned representations or embeddings that capture the temporal and spatial patterns specific to the audio, video, and motion capture modalities. This approach allows leveraging the learned features and representations from the trained CNNs to extract valuable information from the input data across different modalities.

Subsequently, we perform k-fold cross-validation to evaluate the models' generalization and assess potential overfitting by training and validating the model on different subsets (folds) and aggregating the performance metrics to assess generalization. This process involves partitioning the dataset into k equally sized folds, iteratively training the models on $k-1$ folds, and evaluating their performance on the held-out fold. We used 5 folds. This enables us to obtain a more comprehensive understanding of the models' robustness and generalization capabilities.

3.5 Fusion Process

Following the generation of embedding vectors for each modality, we utilize the concatenation fusion technique to merge the vectors and create a unified representation of the input data. The fused embedding vectors are then fed into our CNN model, which consists of two 1D convolutional layers. The first convolutional layer has 16 output channels, a kernel size of 3, a stride of 1, and padding of 1. The second convolutional layer has 32 output channels, a kernel size of 3, a stride of 1, and padding of 1.

In addition to the convolutional layers, our model incorporates batch normalization after each convolutional layer to normalize the activations and enhance training stability. ReLU activation functions are applied after batch normalization to introduce non-linearity. Max pooling operations are performed with a kernel size of 2 and a stride of 2 to reduce the spatial dimensions of the feature maps. The output of the convolutions and pooling layers is then flattened and passed through a fully connected layer. The fully connected layer maps the flattened input to the number of output classes.

4. RESULTS

Our method showed that 1D CNNs can accurately learn spatiotemporal patterns and classify expressive piano performance styles. The unimodal CNN classification results showed that the video CNN achieved an accuracy of 70.57%, the audio CNN attained an accuracy of 86.79%, and the motion CNN achieved an accuracy of 67.92%. In contrast, the multimodal classification model, which fused the information from the video, audio, and motion modalities, achieved an accuracy of 95.29%.

Future work would include comparing single and multiple dimensions in the CNN architecture to see how they affect the classification and learning of the CNN models. Working with other types of architectures, such as Long Short Time Memory (LSTMs) or transformers, would also be interesting, since they effectively handle sequential data, capturing detailed temporal dynamics and complex relationships in music. These models leverage attention mechanisms and long-range dependencies, enabling robust and scalable performance in discerning subtle expressive elements in musical performances. Another interesting approach would be to test audio–video, audio–mocap, or video–mocap combinations and compare their effects on classification performance.

5. DISCUSSION

This study explored multimodal data integration for expressive music performance classification. We incorporated three modalities: video, audio, and motion capture. To comprehensively understand their contributions, we trained separate Convolutional Neural Network (CNN) models for each modality and evaluated their classification performances. The results of our study revealed that the

multimodal approach outperformed individual CNN models. We achieved improved discriminative power and captured diverse aspects of the performances that were not discernible by analyzing each modality independently.

While our study demonstrated the benefits of multimodal data integration for expressive piano performance classification, some limitations should be acknowledged. First, acquiring multimodal data can be challenging due to the scarcity of synchronized multimodal recordings. Collecting a large and diverse dataset with annotated actions across multiple modalities requires significant resources and expertise.

Combining different modalities for fusion requires careful consideration of their compatibility. Not all modalities may effectively complement each other or offer significant synergies. Exploring the ideal combinations of modalities and their fusion techniques is an avenue for further research. Our study utilized a late fusion technique (after feature extraction), but various other fusion methods are available, such as early fusion, hybrid fusion, or attention-based fusion. Exploring different fusion approaches and optimizing the architecture design can improve classification performance.

Furthermore, the effectiveness of the multimodal fusion approach may depend on the specific set of actions and users included in the training data. The ability to generalize to new performance styles and users, especially in real-world scenarios, remains a challenge that warrants further investigation. Lastly, the increased complexity and fusion of multiple modalities may compromise the classification model's interpretability and explainability. Understanding and interpreting how the model arrives at its classification decisions becomes more challenging as the number of fused modalities increases.

6. CONCLUSIONS

Our research aims to lay the foundations for exploring the effectiveness of integrating multimodal approaches and deep learning models in automated expressive music performance recognition during music performances. Through our initial experiments, we have sought to uncover the potential of combining audio, video, and motion data to advance the analysis of expressive music performance.

With a small dataset, we achieved an accuracy of 95%, showcasing our approach's possibilities. By assessing the integration of audio, video, and motion data within a multimodal fusion framework, in conjunction with simple and computationally efficient deep learning models, we have begun to unlock the potential for enhancing the precision and efficacy of expressive music performance recognition. These initial findings provide valuable insights and set the stage for further exploration and refinement.

Moving forward, this research has the potential to inform pedagogy, personalize feedback based on individual abilities, enhance participation and enjoyment in music-making activities, and empower individuals from diverse musical backgrounds. This may open avenues for understanding and analyzing performance nuances, shedding light on a performer's technique, style, and emotional expression. Expressive music performance analysis opens for innovative applications, such as performance feedback systems, personalized music education platforms, as well as immersive and collaborative music experience tools.

Acknowledgments

This project is supported by the Research Council of Norway through project 262762 (RITMO).

7. REFERENCES

[1] A. R. Jensenius, M. M. Wanderley, R. I. Godøy, and M. Leman, "Chapter 2. Musical gestures: concepts and methods in research," 2010.

[2] Sarasúa, B. Caramiaux, A. Tanaka, and M. Ortiz, "Datasets for the Analysis of Expressive Musical Gestures," in *Proceedings of the 4th International Conference on Movement Computing*. London United Kingdom: ACM, Jun. 2017, pp. 1–4. [Online]. Available: https://dl.acm.org/doi/10.1145/3077981.3078032

[3] A. Jensenius, "ACTION – SOUND - Developing Methods and Tools to Study Music-Related Body Movement," Ph.D. dissertation, Jan. 2007.

[4] J. Thompson-Bell, A. Martin, and C. Hobkinson, "'Unusual ingredients': Developing a cross-domain model for multisensory artistic practice linking food and music," *International Journal of Food Design*, vol. 6, no. 2, pp. 233–261, Oct. 2021. [Online]. Available: https://intellectdiscover.com/content/journals/10.1386/ijfd_00032_1

[5] A. R. Jensenius, *Sound actions: conceptualizing musical instruments*. Cambridge: The MIT Press, 2022.

[6] T. Kelkar and A. Jensenius, "Analyzing Free-Hand Sound-Tracings of Melodic Phrases," *Applied Sciences*, vol. 8, Jan. 2018.

[7] A. Sarasua, "Context-aware gesture recognition in classical music conducting," in *Proceedings of the 21st ACM international conference on Multimedia*, ser. MM '13. New York, NY, USA: Association for Computing Machinery, Oct. 2013, pp. 1059–1062. [Online]. Available: https://dl.acm.org/doi/10.1145/2502081.2502216

[8] J. Wu, J. Cheng, C. Zhao, and H. Lu, "Fusing multimodal features for gesture recognition," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, ser. ICMI '13. New York, NY, USA: Association for Computing Machinery, Dec. 2013, pp. 453–460. [Online]. Available: https://dl.acm.org/doi/10.1145/2522848.2532589

[9] X. Chang and L. Peng, "Intelligent Analysis and Classification of Piano Music Gestures with Multimodal Recordings," *Computational Intelligence and Neuroscience*, vol. 2022, p. e8232819, Jun.

2022, publisher: Hindawi. [Online]. Available: https://www.hindawi.com/journals/cin/2022/8232819/

[10] K. Nandakumar, K. W. Wan, S. M. A. Chan, W. Z. T. Ng, J. G. Wang, and W. Y. Yau, "A multi-modal gesture recognition system using audio, video, and skeletal joint data," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, ser. ICMI '13. New York, NY, USA: Association for Computing Machinery, Dec. 2013, pp. 475–482. [Online]. Available: https://doi.org/10.1145/2522848.2532593

[11] "CC BY-NC-SA 4.0 Deed | Attribution-NonCommercial-ShareAlike 4.0 International | Creative Commons." [Online]. Available: https://creativecommons.org/licenses/by-nc-sa/4.0/

[12] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," Austin, Texas, 2015, pp. 18–24. [Online]. Available: https://conference.scipy.org/proceedings/scipy2015/brian_mcfee.html

[13] Y. Jiang, C. Li, and Z. Wang, "Real-time Sound Visualization with Touch OSC: Stimulating Sensibility through rhythm in nature," 2019.

[14] M. Ashraf, F. Abid, I. U. Din, J. Rasheed, M. Yesiltepe, S. F. Yeo, and M. T. Ersoy, "A Hybrid CNN and RNN Variant Model for Music Classification," *Applied Sciences*, vol. 13, no. 3, p. 1476, Jan. 2023, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2076-3417/13/3/1476

[15] N. Bjorck, C. P. Gomes, B. Selman, and K. Q. Weinberger, "Understanding Batch Normalization," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/36072923bfc3cf47745d704feb489480-Abstract.html

[16] H. Wu and X. Gu, "Max-Pooling Dropout for Regularization of Convolutional Neural Networks," in *Neural Information Processing*, ser. Lecture Notes in Computer Science, S. Arik, T. Huang, W. K. Lai, and Q. Liu, Eds. Cham: Springer International Publishing, 2015, pp. 46–54.