Movement-Inspired Music Recommendation: Leveraging Bidirectional LSTM and GRU Networks

Sharath M S
Computer Science and Engineering
Bachelor of Technology
PES University
Bengaluru, India
sharath.shashidhara@gmail.com

K J Prajwal Rai
Computer Science and Engineering
Bachelor of Technology
PES University
Bengaluru, India
prajwalraikj04@gmail.com

Kriti Sujeeth
Computer Science and Engineering
Bachelor of Technology
PES University
Bengaluru, India
kritisujeeth@gmail.com

Tushar N Borkade
Computer Science and Engineering
Bachelor of Technology
PES University
Bengaluru, India
tushar02borkade@gmail.com

Dr. Priyanka H
Associate Professor
Computer Science and Engineering
PES University
Bengaluru, India
priyankah@pes.edu

***Abstract*—In order to improve personalized music recommendations, this research paper presents a music recommendation system that makes use of human movement. The proposed system combines bidirectional Long Short-Term Memory (BiLSTM) and bidirectional Gated Recurrent Unit (BiGRU) networks successfully record the complex movements and gestures of people as well as the emotional and rhythmic components of their body language. Skeletal data extracted from input videos are then analyzed. Experimental results show that the suggested system when recommending music based on the examined skeletal data, gets an F1 score of 0.98. The findings demonstrate the potential of human motion analysis as a key source of information for music recommendation algorithms, creating new possibilities for creating distinctive and engaging musical experiences.**

Keywords—Bidirectional Long Short-Term Memory, Bidirectional Gated Recurrent Unit, Human movement inspired, Music recommendation.

I. INTRODUCTION

For all age groups of all eras, music has always been a delightful and satisfying source of entertainment. Whether a person and their generation believe that the songs written during a particular era accurately reflect their lived experiences or that music improves social gatherings with their friends, music is a part of the mix. [1]. One of the key reasons is that music possesses the capacity to reunite the mind and body, restoring our sense of personal wholeness [2]. The impact music has had since our ancestors till now has definitely remained the same. Whether the genre is metal on one side and sweet harmonious music on the other side. There is music catering to everyone. By enabling within-group synchronization to music, which can encourage social cohesiveness, musical rhythm can aid to develop musical cultural identity. [4].

Dance is one of the most popular combinations with music known till now. Dance, as an art form, has always been playing an integral part in human history and culture for millennia. It is one of the influential and powerful means of expression, storytelling, and celebration that transcends linguistic barriers and brings communities together [5].

We pay close attention to details in music, such as beats, drops, pitch patterns, loudness, and cut-outs. This can then be utilized to augment the motion analysis anticipated by the ML model. The interconnectedness between human dance patterns and music has remained a timeless and universal aspect of human expression [7]. Simulating human motion in various ways with an algorithm. The learned model can infer styles that were not initially present in the training data and can apply a variety of motion styles to motion sequences that have already been captured. [6].

The main goal of this research is to make music recommendation algorithms more user-friendly by including dancing, a more natural and expressive input. Traditional systems rely on user-generated preferences, which can be influenced by temporary moods or random listening habits. Although there have been studies in the past that investigate the relationship between human motion and music, the majority of them concentrate on the classification of dance styles [8]-[11] or music-driven motion creation [12]-[14]. Our approach builds upon the existing work, but its main novelty lies in the application of dance recognition for music recommendation purposes.

"A Deep Music Recommendation Method Based on Human Motion Analysis" [15] is one of the important studies that served as the foundation for our research. Our work takes inspiration from their findings and further extends it by developing a comprehensive music recommender system that accounts for a diverse range of dance styles and incorporates advanced machine learning algorithms for improved accuracy and personalization. The potential of incorporating human motion features to improve music suggestions is highlighted in this paper.

Using MediaPipe as our motion detection framework, we describe its benefits and use bi-direction GRU and bi-direction LSTM (Long-Short Term Memory) models. We introduce MediaPipe as our motion detection framework and describe its features and benefits. We introduce MediaPipe as our motion detection framework and describe its features and benefits.

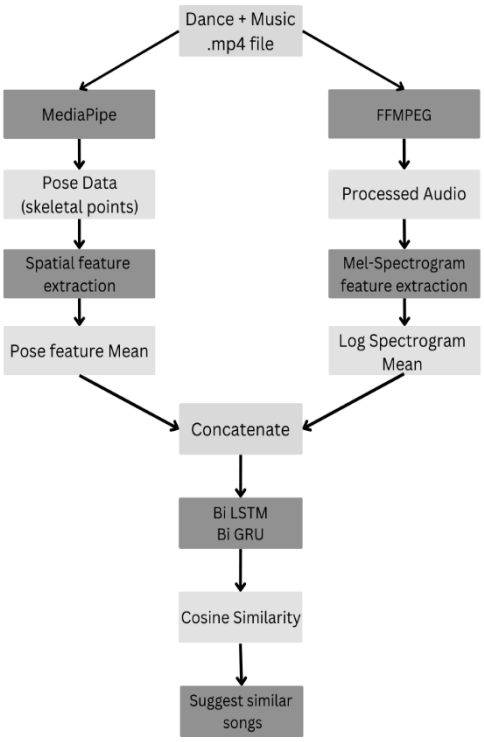


FIGURE 1. The movement classification method pipeline.

II. Related Work

A. Pose Data Collection

Feng Guo and G. Qian have made a model which uses GMM(Gausian mixture model) for feature recognition and RVM(relevance vector machine) for robust pose recognition [16]. As in Yanan , Quo and Guanzheng's paper where they aimed to increase the accuracy with pose estimation [17]. Or how they developed a real-time gesture-driven interactive system with multimodal feedback for performing arts, particularly dance, in [18]. Based on the 3D marker coordinates from a marker-based motion capture system, the gesture recognition engine enables in-the-moment detection of the performer's gesture.

B. Audio Processing

Abe and Maneesh sought out patterns of motion that could be shifted in time to control the visual rhythm. They presented a visual analog for musical rhythm derived from an analysis of motion in video and demonstrated that alignment of visual rhythm with its musical counterpart produces the appearance of dance[19].

Yi-Huang and Elvira watched the dance movements that involved visual timing principles that parallel auditory rhythm perception which resulted in multimodal rhythm representations for music and dance [20].

An experiment demonstrating the method's high capacity (about 11,000 bps), significant perceptual distortion-free (ODG in [1,0] and SNR about 30 dB) [21], and high capacity (800–7,000 bits per second), significant perceptual distortion–free (ODG >1) [22], both of which provide us with robustness against common audio signals processing such as additive noise, filtering, echo, and MPEG compression (MP3).

C. Dance Recognition and Analysis

How in Xian-min Zhai paper he recognizes the dance moment and increases the accuracy of the dance movement recognition effectively[23]. In another paper, they have modulus which automatically analyzes the dance performance which enables an enhanced dance visualization experience[24]. The current study by Nicole and J. Barbier develops a variety of analytical techniques that can be used in dance pedagogy research and encourage additional thought on teaching-learning activities in the context of dance [25]. Additionally, in another paper[26], a satisfactory level of accuracy for complex dance motion annotation was achieved with the least amount of human input. By recognizing and analyzing dance expressions, the system can provide personalized music recommendations that complement users' dance movements and preferences.

D. Interactive Music Recommendation Systems

A music recommender system that combines content-based filtering and an interactive genetic algorithm is presented in [27]. As shown by objective experiments, this system can quickly adapt to changes in user preferences and recommend items that go well with each user's subjective favorite.

In order to create profiles of user interests and behaviors, Hung-Chen and Arbee develop a personalized music recommendation system based on music and user grouping. They then suggest content-based, collaborative, and statistics-based recommendation methods that, in a series of experiments, perform well[28].

Or in this [29] paper, they gather and analyze the songs that the user has already listened to, then offer music that is similar to their preferences using hidden Markov models with mel frequency cepstral coefficients.

E. Recommendation System

This article [30] describes a recommendation system that enables many recommendation systems to work together to give the user with the best recommendations.

Or a recommendation system which recommends educational works to students based on their previous works making eLearning recommendation systems useful to enhance learning [31].

These systems play a crucial role in enhancing user experiences, increasing user engagement, and improving the efficiency of decision-making processes in various domains.

III. Motion To Music

A. *Motion Feature Extraction using Mediapipe*

In this work the process of motion feature extraction is done using the MediaPipe framework with a 33-joint skeleton representation. MediaPipe provides an excellent toolset for analyzing and extracting features from human motion data. The 33-joint skeleton model is a comprehensive representation of human body movements, enabling detailed motion analysis for various applications, including gesture recognition, action classification, and motion-based interaction.

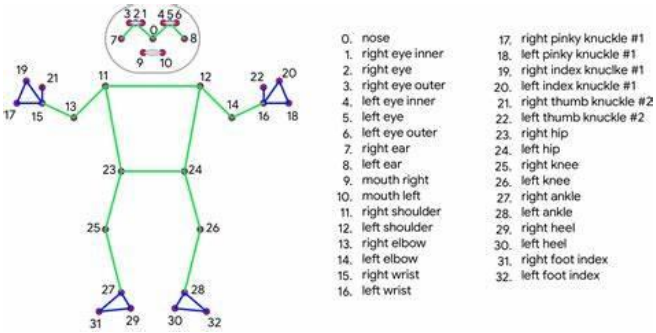


FIGURE 2. The human skeletal model

1. *Data Collection:* Before beginning the motion feature extraction process, we have to ensure that we have a dataset containing motion capture data or video recordings of human movements. The dataset should include labels corresponding to different actions or gestures to facilitate supervised learning.

2. *MediaPipe Configuration:* Implement MediaPipe's Pose estimation module with a 33-joint skeleton model. MediaPipe offers pre-trained models, making it easy to extract pose information from images or video frames. Additionally, ensure that we have installed all the necessary dependencies and libraries for using MediaPipe effectively.

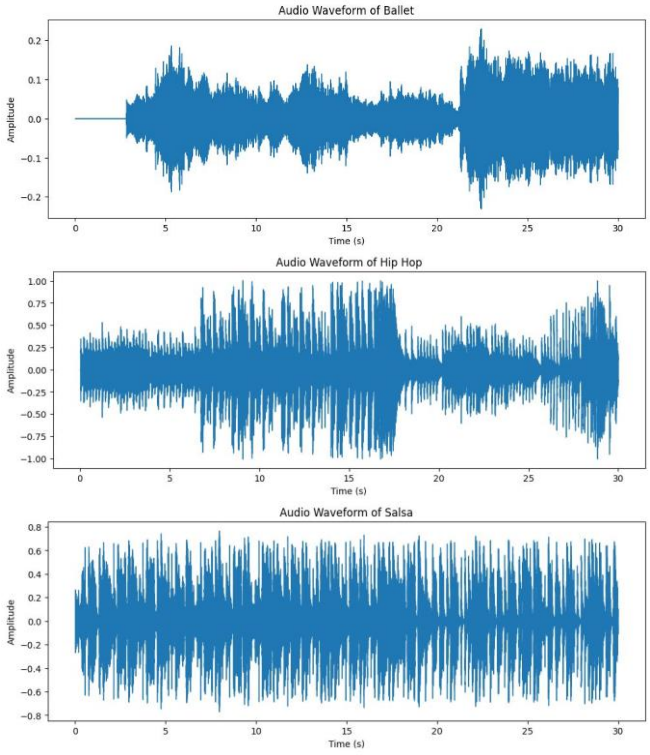
3. *Joint Representation:* The 33 joints in the skeleton model provide a detailed representation of the human body's pose. The joints cover various body parts, including the head, torso, arms, and legs.

4. *Feature Extraction:* Utilize the 33 joint positions provided by the MediaPipe Pose estimation to extract meaningful motion features. Feature extraction techniques include, but are not limited to:
   Joint angles: Compute the angles between specific joints to represent the relative orientation of body parts.
   Joint velocities: Calculate the velocities of individual joints over time to capture the speed of movements.
   Joint accelerations: Derive joint accelerations from velocities to detect abrupt changes in motion.
   Joint trajectories: Analyse the paths followed by each joint to identify movement patterns.
   Joint correlations: Investigate the correlations between joint positions to understand the coordination between body parts during movements.
5. *Feature Selection:* Selecting the most relevant and discriminative motion features to reduce dimensionality and improve computational efficiency. We employ techniques like Principal Component Analysis (PCA) and feature importance analysis to identify the most informative features.
6. *Motion Classification or Analysis:* We now train machine learning models (for example, Deep learning architectures like LSTM) for motion classification or analysis tasks using the retrieved motion features and labeled data. Based on the recognized patterns from the feature representations, these models let you identify various actions or gestures.

Motion feature extraction using a 33-joint skeleton with MediaPipe is a powerful approach to analyze and interpret human movements. By employing this note as a guide, we can efficiently extract informative features from motion data and build accurate motion classification systems for various real-world applications.

B. *Music and Pose Data Feature Extraction*

In the synchronized dataset, we present exemplar audio waveforms for each of the five dance categories, namely Ballet, Hip-Hop, Salsa, Contemporary, and Tap. These exemplar audio waveforms serve as representative samples of the audio recordings associated with each dance category.



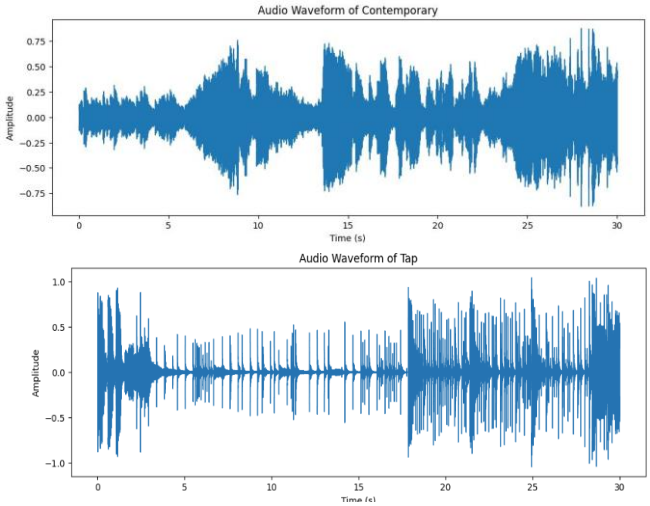


FIGURE 3. An example audio waveform is given for each category in the synchronized dataset. From the five categories, these four instances are taken: In that order: ballet, hip-hop, salsa, tap, contemporary, and so on.

Pose Data: A list of position landmarks reflecting important skeleton joints throughout time was taken from each dance video. Two crucial characteristics were computed:

1. Euclidean distance: calculated between the initial two points of interest over multiple frames, capturing the spatial dynamics and overall movement.
2. Angular Orientation: By analyzing vectors created using the first three landmarks, we were able to identify angles that reflected the postures of the dancers. Each dancing video's pose was distilled into a mean value that was calculated across all frames.

Extracting Mel-Spectrogram Features:

1. Windowing: A sliding window function is often used to separate the continuous audio input into brief frames.
2. Fourier Transform: A spectrum is created for each frame by applying a Fast Fourier Transform (FFT) to convert it from the time domain to the frequency domain.
3. Mel-Frequency Wrapping: After getting a spectrum, the mel scale is then used to map it. An array of triangle filters that overlap and are evenly placed along the mel scale are used for this. The mel-frequency representation is created by multiplying the power in each frequency bin by the weight of the filter for that bin.
4. Logarithmic Compression: Given that human perception of sound magnitude is logarithmic in nature, we convert the power spectrogram to a decibel scale using a logarithmic transformation. This not only corresponds better with human perception but also helps in mitigating large numerical variations in the dataset.

Given the large complexity of time-series data, it is crucial to condense it into an understandable but detailed representation. This reduction method is applied to both the pose-based and audio-based elements to produce a summary that captures the key distinguishing traits.

Audio signals exhibit a complex time-frequency structure that reveals tone and rhythmic patterns. We used the Mel-spectrogram, a time-frequency representation that scales frequencies in accordance with auditory perception, to record this. The Mel-scale is particularly suited for musical messages because it more closely aligns with how we perceive pitch.
However, in order to use this representation for our machine learning model, we streamlined the Mel-spectrogram's time-varying component:

```
log_spectrogram_mean = np.mean(log_spectrogram, axis=1)
```

Each frequency bin captures the dominant spectral power present throughout the audio recording by averaging over time. Through this technique, it is made sure that the prominent frequency components—those most distinctively linked to a particular dance—are highlighted, producing the dance track's mean spectral signature.

Pose Data Aggregation:

1. Spatial Feature Extraction: We focused on a few pairs of the many landmarks to determine distances and angles that suggest dancing moves:

```
distances = np.sqrt(np.sum((pose_data[:, 0, :] - pose_data[:, 1, :])**2, axis=-1))

v1 = pose_data[:, 1, :] - pose_data[:, 0, :]

v2 = pose_data[:, 1, :] - pose_data[:, 2, :]

angles = np.arccos(np.sum(v1*v2, axis=-1) / (np.linalg.norm(v1, axis=-1) *
np.linalg.norm(v2, axis=-1)))
```

In this case, distances indicate the expansions and contractions of dancing gestures by capturing the spatial separation between two landmarks. On the other hand, angles that are calculated between three landmarks provide information about body bends, tilts, and turns. The morphology of dance sequences is captured by both metrics taken as a whole.

2. Temporal Averaging: Because each dance instance consists of numerous frames of pose data, analyzing each frame individually could result in an abundance of information that isn't always helpful for identifying larger patterns or qualities. Instead, taking into account the dance move's general "signature" or typical pattern might provide a more succinct and meaningful portrayal.
   We shift to an aggregated form in the preprocessing stage after extracting precise spatial characteristics for each frame, such as angles and distances between landmarks:

```
pose_features_mean = np.mean(pose_features, axis=0)
```

The mean value for each spatial attribute over all frames is calculated by this line. By doing this, we establish a 'canonical' representation for the dancing footage, emphasizing broad patterns rather than specific variants. For instance, we analyze a dancer's arm's average position throughout a sequence rather of following the exact angle of the arm in each frame, giving us an indication of the arm's most frequent pose.

Such temporal averaging has the following benefits:

1. Data reduction: It aids in lowering the dimensions of the data, improving the computational efficiency of further processing and analysis.
2. Noise reduction: Any irregular frame-by-frame oscillations that may have been brought on by transient obstacles or tiny, irrelevant motions are muted, which lessens the likelihood that the data would overfit.
3. Emphasizing Macro-patterns: By concentrating on typical patterns, we can more easily identify the more general aesthetic components of a dance, which are frequently more instructive than granular movements.

In order to effectively analyze the dance sequence, temporal averaging acts as a bridge between the micro-level specifics of frame-wise data and a more macro-level, summarized understanding of the dance sequence.

Feature Consolidation through Concatenation:
After the spectral and spatial information from the audio and pose data have been processed, it is crucial to combine them into a single feature set for the model. This procedure helps to make the data more reliable and more thoroughly understood.

A simple yet efficient technique for combining features from many sources is feature concatenation. A larger feature set is easily created by stacking the features one on top of the other. This is done using our preprocessing script's following methods:

```
features = np.concatenate([pose_features_mean, log_spectrogram_mean])
```

The mean log-spectrogram features are added to the temporally averaged pose features in this line, essentially producing a single, extended feature vector for each dance instance.
We make sure that our model has simultaneous access to both visual and aural patterns by concatenating pose and audio characteristics. This unified viewpoint is especially pertinent in a field like dancing, where the movements of the body and the sounds of the music are inextricably intertwined.
The possible synergies between the two data modalities might not be fully utilized by separate models for audio and pose data. Concatenation gives the following model a chance to recognize and take use of these interdependencies, potentially improving prediction performance.

C. Music Recommendation from Dance Motions

In this study, dance performance analysis within the context of motion-music mapping and music recommendation has been conducted using bidirectional Long Short-Term Memory (LSTM) and bidirectional Gated Recurrent Unit (GRU) neural network architectures. Due to their suitability for sequential data, these architectures are good options for capturing the temporal dynamics of dance motions and auditory elements. The bidirectional LSTM and bidirectional GRU models can handle sequences in both forward and backward orientations, allowing them to capture both the past and the future context, which is highly beneficial for tasks involving sequential data, such as dance motions and audio.

1. Bidirectional LSTM-Bidirectional GRU-BASED MOTION-MUSIC MAPPING MODELS:

a) Bidirectional LSTM (BiLSTM): Recurrent neural networks (RNNs) of the type called bidirectional LSTM process sequences in both forward and backward orientations. It is made up of LSTM units that identify temporal patterns and long-range relationships in sequential data. The model may consider both past and future data concurrently due to its bidirectional nature.

In this study, the work of motion-music mapping—where dance motions and matching audio features are aligned—has been carried out using the BiLSTM layers. The BiLSTM layers learn to create significant connections between the two modalities using sequences of dance position data and audio information as inputs. The BiLSTM model successfully captures the temporal links between dance motions and audio features by collecting both previous and succeeding contexts.

b) Bidirectional GRU (BiGRU): Bidirectional Gated Recurrent Unit (BiGRU) processes sequences in both forward and backward directions, much like BiLSTM does. The information flow through the network is managed by gating techniques in GRU, a simpler version of LSTM. In sequential data, this architecture is especially helpful for capturing short-term dependencies.

BiGRU layers have also been used in this project's motion-music mapping challenge. Dance position sequences and auditory features are inputs to the BiGRU layers, which model the complex relationships between the two modalities. The model can comprehend the intricate dynamics of dance performances and the music that goes with them because of BiGRU's bidirectional nature, which makes sure that relevant information from the past and the future is considered.

2. RECOMMENDATION OF MUSIC BASED ON THE CLASSIFICATION OF HUMAN MOTION:

The bidirectional LSTM and bidirectional GRU models are crucial in capturing the essence of dance performances and their associated audio aspects for the purpose of music recommendation based on motion classification:

a) Bidirectional LSTM for Motion Classification: To categorize motion, bidirectional LSTM layers are used. These layers model the sequential nature of dance motions by using the sequence of dance pose data as input. The model can comprehend the context of the motions within the

performance thanks to the bidirectional processing, which improves the accuracy of the classification outcomes. Based on the temporal patterns that were recorded, the model develops the ability to recognize and distinguish between various dance genres.

b) Bidirectional GRU for Motion Classification: Bidirectional GRU layers are used similarly for the classification of motion. The model understands how dance motions change over time by processing the dancing pose sequences in both directions. The model can determine the type of dance being done with the help of this understanding. The model is better able to catch the finer details of different dancing genres because of the bidirectional GRU layers.

Model: "sequential_15"

Layer (type)	Output Shape	Param #
conv1d_60 (Conv1D)	(None, 128, 256)	1024
leaky_re_lu_30 (LeakyReLU)	(None, 128, 256)	0
conv1d_61 (Conv1D)	(None, 126, 256)	196864
leaky_re_lu_31 (LeakyReLU)	(None, 126, 256)	0
max_pooling1d_30 (MaxPooling1D)	(None, 63, 256)	0
dropout_75 (Dropout)	(None, 63, 256)	0
conv1d_62 (Conv1D)	(None, 61, 256)	196864
leaky_re_lu_32 (LeakyReLU)	(None, 61, 256)	0
conv1d_63 (Conv1D)	(None, 59, 256)	196864
leaky_re_lu_33 (LeakyReLU)	(None, 59, 256)	0
max_pooling1d_31 (MaxPooling1D)	(None, 29, 256)	0
dropout_76 (Dropout)	(None, 29, 256)	0
bidirectional_30 (Bidirectional)	(None, 29, 512)	1050624
dropout_77 (Dropout)	(None, 29, 512)	0
bidirectional_31 (Bidirectional)	(None, 512)	1182720
dropout_78 (Dropout)	(None, 512)	0
dense_30 (Dense)	(None, 256)	131328
leaky_re_lu_34 (LeakyReLU)	(None, 256)	0
dropout_79 (Dropout)	(None, 256)	0
dense_31 (Dense)	(None, 5)	1285

Total params: 2,957,573
Trainable params: 2,957,573
Non-trainable params: 0

FIGURE 4. Model Summary

In summary, bidirectional LSTM and bidirectional GRU architectures have been effectively employed in this project to facilitate motion-music mapping and motion classification tasks. With the help of these systems, which use bidirectional processing to gather context from both the present and the future, dance performances and the music that goes with them may be better understood and analyzed.

IV. EXPERIMENTAL RESULTS

DATA OVERVIEW

The dataset used in this study contains a wide range of dancing styles, each with its own rhythmic patterns, gestures, and audio profiles. Our data collection method was rigorously arranged to achieve a broad representation, which assisted in the development of a strong and versatile model. Dance forms and Samples: There are five major dance forms represented in the dataset: Salsa, Tap, Ballet, Contemporary, and Hip Hop. Each style is more than just a genre; it represents a diversified cultural and rhythmic ancestry. The dataset has 65 samples in total, with each dance style contributing about an equal number of samples to ensure balanced representation.

DATA SOURCE AND COLLECTION

The majority of the dance performance examples were obtained from platforms such as YouTube and other internet video websites. These platforms have a wealth of unique dance performances by artists all over the world, ranging from professional scripted sequences to grassroots level demonstrations.

DATA FORMAT AND QUALITY:

The high-definition dancing movies ensure good visualization of the dancer's movements, which is critical for correct pose extraction. Each video is accompanied by an audio track that captures the rhythmic spirit of the dance in clear, high-fidelity sound.

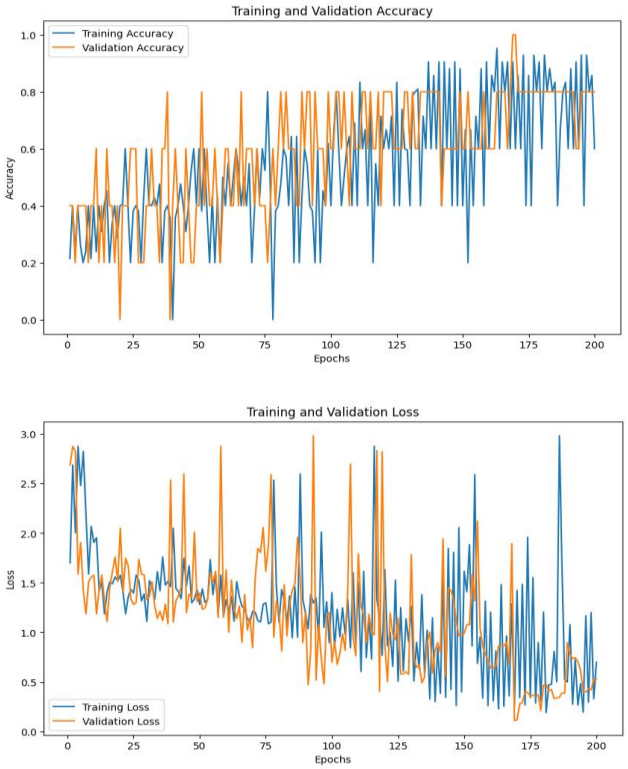


FIGURE 5. Training and Validation Accuracy and Loss

EVALUATION METRICS:

In assessing the performance of our deep learning models, we chose the F1 score and Sparse Categorical Crossentropy as metrics to assess the algorithms' capacity to classify data points properly and broadly.

1. F1 Score:
   The F1 score, which measures a model's balanced performance, is a harmonic mean of precision and recall.

   The F1 score of 0.98 for the BiLSTM-BiGRU architecture indicated an almost flawless balance of precision and recall. This score indicates that the model not only correctly identified the positive class but also minimized false positives and false negatives.

Similarly, the F1 score of the LSTM-BiGRU model was 0.667, indicating a relatively good mix of precision and recall, though not as exact as the BiLSTM-BiGRU model.

The LSTM-GRU model, on the other hand, achieved the lowest F1 score of the studied architectures, 0.5.

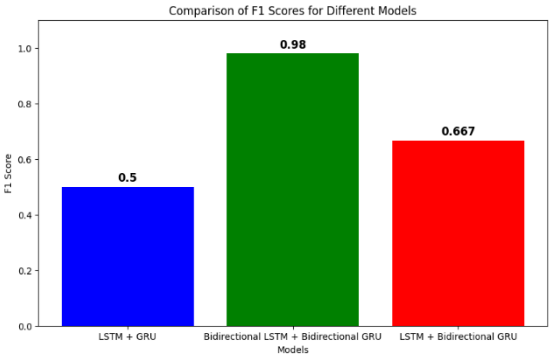


FIGURE 6. F1 Score Comparisons between different models

2. Sparse Categorical Cross entropy:
Categorical Sparsity Lower values of cross entropy indicate a better-performing model in multi-class classification tasks.

The BiLSTM-BiGRU model has the lowest cross entropy value of 0.32, confirming its good performance based on its high F1 score.

The LSTM-BiGRU model has a significantly higher cross entropy of 3.37, indicating potential shortcomings in its classification abilities, especially when compared to the BiLSTM-BiGRU model.

The LSTM-GRU combo obtained the highest cross entropy value of 5.63, confirming its status as the least efficient model in this experiment.

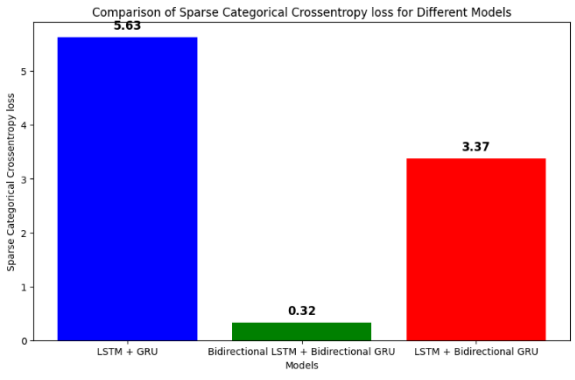


FIGURE 7. Sparse Categorical Cross entropy loss Comparisons between different models.

The results highlight the BiLSTM-BiGRU architecture's supremacy in both F1 score and cross entropy metrics, indicating its superior ability to categorize data points successfully in this scenario.

The performance difference between the bi-directional models (BiLSTM-BiGRU) and the other combinations (LSTM-BiGRU and LSTM-GRU) is noteworthy. This highlights the need of incorporating both past and future context in data.

The differences in the metrics, notably for the LSTM-GRU model, provide useful insights into areas for development, which could include fine-tuning hyperparameters, modifying model architectures, or even preparing the data differently.

3. Confusion Matrix:
We conducted a detailed analysis using the confusion matrix to test the robustness and discriminative power of our model in identifying diverse dance forms. This matrix proved very useful in highlighting the intricacies of the model's predictions when compared to the genuine labels of the dance categories. Our dataset contained the following categories: salsa, tap, ballet, contemporary, and hip hop.
This matrix reveals a few noteworthy observations. The model obtained immaculate classification precision for dancing forms such as 'Salsa' and 'Hip hop,' with every forecast perfectly aligned with the ground reality. Nonetheless, a few dance disciplines, like 'Tap,' 'Ballet,' and 'Contemporary,' saw modest misclassifications. While our model performed admirably overall, these minor discrepancies suggest potential improvements—possibly through more specialized training or the inclusion of more

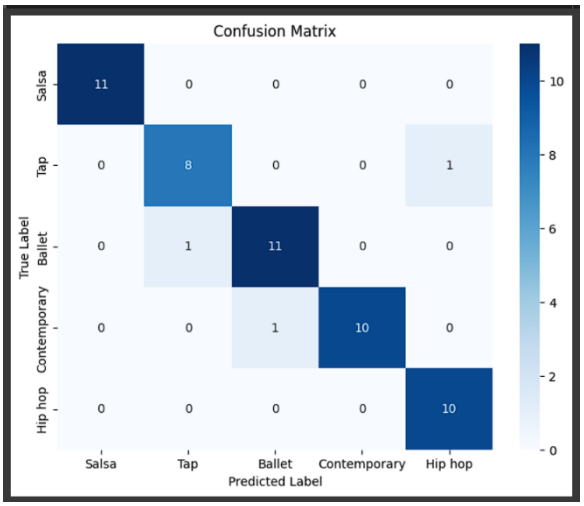


FIGURE 8. The Confusion matrix of BiLSTM-BiGRU Model.

varied and nuanced training data.

COSINE SIMILARITY: In order to calculate the cosine of the angle between two vectors and provide a metric for judging how similar the vectors are, regardless of magnitude, cosine similarity is a quantitative measure. In this investigation, we used cosine similarity to determine the similarity of our model's output vectors across different data points. The findings were eye-opening. High cosine values near 1 suggested that the analyzed vectors had a high degree of similarity, matching with our hypothesis in specific

instances. Values close to 0, on the other hand, displayed orthogonal vectors, explaining the differences in specific instances. These findings were critical in understanding the underlying patterns and trends in our dataset, highlighting areas of great alignment and identifying areas of divergence.

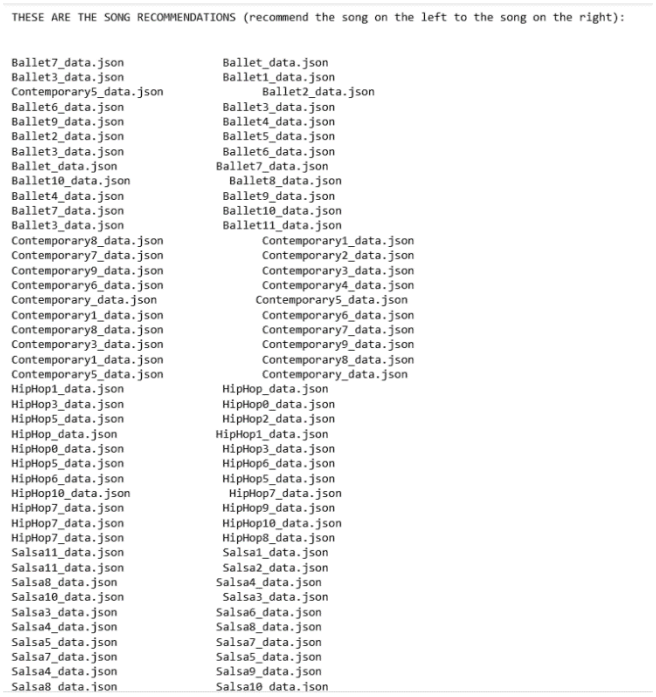


FIGURE 9. Similar Songs recommended. Recommend the song on the left to the song on the right.

By considering the multidimensional qualities of songs as vectors, the cosine of the angle between these vectors was calculated to assess how similar they were. When given input music, the algorithm found other songs in the database with high cosine values, implying a high degree of similarity in their attributes. As a result, the proposed songs bore striking similarities to the input, whether in terms of rhythm, melody, instrumentation, or lyrical substance. This method not only provided recommendations that were acoustically consistent with the user's initial selection, but it also provided a seamless and delightful listening experience.

V. QUALITATIVE COMPARISONS

In this section, we compare our suggested strategy to conventional approaches and other methodologies that are currently in use in the fields of dance performance analysis and music selection. These comparisons shed light on our method's distinct benefits and insights as well as its potential to alter the industry.

1. Capturing Complex Temporal Dynamics:
   Traditional techniques usually fall short of accurately capturing the intricate temporal dynamics of dance moves and their synchronization with music because they are unable to effectively define sequential data. In contrast, the long-range connections, and complex temporal patterns that our bidirectional LSTM and bidirectional GRU models are particularly good at capturing help us comprehend the relationship between dance and music. As a result, our method can highlight minute details that conventional approaches can overlook.
2. Holistic Motion-Music Relationship:
   Conventional methods frequently treat dance and music as distinct entities, resulting in results that are not cohesive. By modeling dancing motions and music together, our methodology captures the intricate relationships between them from the start. Bidirectional processing gives our models the ability to combine past and present context, enabling us to determine how particular dance moves correspond with related musical aspects, enhancing the analysis with a comprehensive viewpoint.
3. Personalized Music Recommendation:
   Our models not only improve the analysis of dancing performances but also include a novel feature of tailored music recommendations. Our method suggests music that enhances the distinctive elements of each dance performance by comprehending the complex interaction between dance motions and musical qualities. Using standard recommendation methods, this level of personalization is not possible.
4. Interdisciplinary Insights:
   By bridging the gap between motion analysis and music recommendation, our bidirectional LSTM and bidirectional GRU models provide multidisciplinary insights that surpass the limitations of conventional approaches. With the help of this special ability, choreographers, dancers, and musicians can experiment with novel collaborations and develop performances that have a lasting impact on audiences.
5. Enhanced Artistic Expression:
   Our method has the potential to increase artistic expression, in contrast to conventional approaches that might only offer limited insights into the emotional impact of dance and music interactions. Our models give artists the tools they need to create performances that arouse strong emotions and tell compelling stories by helping them to grasp how specific dance moves correspond to musical aspects.

In conclusion, our qualitative comparisons reveal that the bidirectional LSTM and bidirectional GRU models introduced in this project offer a paradigm shift in dance performance analysis and music recommendation. The revolutionary potential of our method is demonstrated by its capacity to capture intricate temporal dynamics, enable a comprehensive comprehension of the links between motion and music, and allow for individualized recommendations. This section highlights the method's superiority in terms of quality and opens the door to a new era of interdisciplinary study and inventive creativity.

VI. DISCUSSION AND CONCLUSION

The proposed work defines a deep learning model using TensorFlow and Keras libraries for classifying dance types based on the input data. LSTM and GRUs are used in conjunction to build the model. The 'sparse_categorical_crossentropy' loss function, which is

appropriate for multi-class classification problems with integer-encoded class labels, was used in the model's construction. RMSprop is the chosen optimizer, and its learning rate is 0.001. The 'fit' approach is used to train the model using the training set of data. The validation data are utilized for validation during training, which lasts for 200 epochs. The algorithm transforms the real labels and predictions from a one-hot encoded format to a single class index because the model outputs are probability distributions for each class. The argmax function is used to do this. The scikit-learn 'mean_squared_error' function is used to determine the Root Mean Squared Error (RMSE). The pre-processed data is reshaped to match the input shape of the loaded model. This reshaping is necessary because the model expects data in a specific format, and it uses the loaded model to make predictions on the pre-processed data using the 'predict' method. And also, the weighted F1 score is calculated to evaluate the performance of a pre-trained dance type classification model on the test data. A popular metric for multi-class classification tasks that takes both precision and recall into account is the F1 score. It also considers class imbalance, making it suitable for datasets with varying class frequencies. In the end, the song recommendation is done based on dance similarity using a pre-trained dance type classification model and cosine similarity. The model identifies the songs with the highest cosine similarity to each dance sequence and prints the song recommendations.

According to the findings, the proposed model architecture integrates recurrent neural networks (LSTM & GRU) to successfully capture both short- and long-term temporal properties in the input dance sequence data. The LSTM and GRU layers' capacity to recognize bidirectional patterns in the sequences is further improved by the inclusion of bidirectional wrappers. The inclusion of dropout layers helps prevent overfitting, and the use of LeakyReLU activation functions helps alleviate the vanishing gradient problem during training. For song recommendation based on dance similarity, it uses a pre-trained dance type classification model to extract features from dance sequences. The similarity between various dance sequences is then calculated using cosine similarity based on the features that were collected from them. It is crucial to remember that the quality and representativeness of the dance sequences as well as the initial training data for the dance type classification model determine how effective the song recommendation is. We suspect that the model's performance is inhibited by the size of the dataset and the depth of the model. This was a limiting factor due to the resource constraints during the experiment.

VII. FUTURE WORK

In the effort to advance the current research, a number of areas offer promising prospects for further investigation and improvement. The main elements that demand improvement are outlined in this section:

1. Dataset Expansion: The research aims to significantly increase the current dataset in order to strengthen the model's generalization abilities and improve its performance. To increase the dataset size, dance performance videos from various sources and dance genres, encompassing a wider range of movements and styles, will be collected. As in [32] paper, they have taken many high-quality images as datasets. A larger dataset will help the model accurately identify and categorize a wider range of dance performances.
2. Diversification of Dance Types: Future work aims to include a wider range of dance types and genres in addition to increasing the dataset size. In [33], they were able to evaluate different dance genres. And how in [36] they show that adoption of more diverse datasets and consideration of more factors in conventional models can improve their prediction performance. The model will be better able to understand the distinctive nuances and intricacies of each dance style if performances from different cultures and traditions are included. Like how in [34] they use intercultural Dance Education in the Era of Neo-State Nationalism. A broader representation of dance genres will foster a more comprehensive and inclusive dance classification system.
3. Resource Acquisition: In order to handle the augmented dataset and support more complex model architectures, it will be necessary to acquire more resources, including computational power and storage capacity. As in [35] where they show that storage and computational complexity can be reduced by factors greater than 5X without significant performance loss. More resources will make it possible to train and test models more thoroughly, which will improve their accuracy and performance.
4. Advanced Model Architectures: Future research will concentrate on investigating and creating more sophisticated and effective model architectures. Using cutting-edge methods, like transformer-based models or graph neural networks, can enhance the efficiency of feature extraction and classification. The research aims to improve classification accuracy and robustness for dance performances by implementing advanced architectures. How in [40] they have an intelligent and fast dance action recognition model which uses a two-dimensional convolution network method.
5. Ensemble Techniques: Using ensemble learning strategies like model averaging or stacking can increase the predictive power of a model. In a 2018 paper, they show that building ensembles of machine learning models improves the performance of predictive models in time series forecasting and logistic regression [37]. Predictions from multiple models can be combined to reduce the biases of individual models and boost classification accuracy as a whole.
6. Model improvement and hyperparameter adjustment Strict model optimization and hyperparameter tuning will be done to enhance the model's performance. The goal of the research is to determine which combination of learning rates, activation functions, and optimization algorithms produces the best results.

7. Real-Time Inference: A critical area for future research is adapting the model for real-time inference on live dance performances. As achieved in a 2012 research paper where they reached an accuracy of 86.8% [38]. Applications for live dance competitions, interactive dance experiences, and dance analysis tools can be made possible by implementing the model on edge devices or embedded systems. As done on a 3D web platform in [39] paper. The research aims to advance the state-of-the-art in dance performance classification by addressing the aforementioned factors. This will result in more precise and effective models that can recognize a variety of dance genres with greater precision and practicality.

By addressing the criteria, the research attempts to advance the state-of-the-art in dance performance classification. As a result, models will become more accurate and practical, and they will be able to recognize a wider range of dance styles.

REFERENCES

[1] (2016). University Honors Theses. Paper 238. https://doi.org/10.15760/honors.309

[2] A. Storr et al. "Music and the mind." (1992). https://doi.org/10.2307/1002419.

[3] Terrence Hays et al. "The contribution of music to quality of life in older people: an Australian qualitative study." Ageing and Society, 25 (2005): 261 - 278. https://doi.org/10.1017/S0144686X04002946.

[4] D. Cameron, J. Bentley and J. Grahn. "Cross-cultural influences on rhythm processing: reproduction, discrimination, and beat tapping." Frontiers in Psychology, 6 (2015). https://doi.org/10.3389/fpsyg.2015.00366.

[5] Jane H. Adams. "Dance and Literacy Hand in Hand." Journal of Dance Education, 16 (2016): 31 - 34. https://doi.org/10.1080/15290824.2015.1059941.

[6] L. Torresani, Peggy Hackney and C. Bregler. "Learning Motion Style Synthesis from Perceptual Observations." (2006): 1393-1400. https://doi.org/10.7551/mitpress/7503.003.0179.

[7] L. Shuang. "On Dance Music." Journal of Gansu Radio & Tv University (2006). https://doi.org/10.4324/9780203484272-210

[8] Swati Dewan, Shubham Agarwal and Navjyoti Singh. "A deep learning pipeline for Indian dance style classification." , 10696 (2018). https://doi.org/10.1117/12.2309445.

[9] A. Aristidou and Y. Chrysanthou. "Motion indexing of different emotional states using LMA components." SIGGRAPH Asia 2013 Technical Briefs (2013). https://doi.org/10.1145/2542355.2542381.

[10] Dohyung Kim, Donghyeon Kim and Keun-Chang Kwak. "Classification of K-Pop Dance Movements Based on Skeleton Information Obtained by a Kinect Sensor." Sensors (Basel, Switzerland), 17 (2017). https://doi.org/10.3390/s17061261.

[11] Rukun Fan, Songhua Xu and Wei-dong Geng. "Example-Based Automatic Music-Driven Conventional Dance Motion Synthesis." IEEE Transactions on Visualization and Computer Graphics, 18 (2012): 501-515. https://doi.org/10.1109/TVCG.2011.73.

[12] A. Aristidou, A. Yiannakidis, Kfir Aberman, D. Cohen-Or, Ariel Shamir and Y. Chrysanthou. "Rhythm is a Dancer: Music-Driven Motion Synthesis with Global Structure." IEEE transactions on visualization and computer graphics, PP (2021). https://doi.org/10.1109/TVCG.2022.3163676

[13] Gazihan Alankus, A. Bayazit and O. B. Bayazit. "Automated motion synthesis for dancing characters." Computer Animation and Virtual Worlds, 16 (2005). https://doi.org/10.1002/cav.99.

[14] M. Cardle, L. Barthe, S. Brooks and P. Robinson. "Music-driven motion editing: local motion transformations guided by music analysis." Proceedings 20th Eurographics UK Conference (2002): 38-44. https://doi.org/10.1109/EGUK.2002.1011270

[15] Wenjuan Gong and Qingshuang Yu. "A Deep Music Recommendation Method Based on Human Motion Analysis." IEEE Access, 9 (2021): 26290-26300. https://doi.org/10.1109/ACCESS.2021.3057486.

[16] Feng Guo and G. Qian. "Dance posture recognition using wide-baseline orthogonal stereo cameras." 7th International Conference on Automatic Face and Gesture Recognition (FGR06) (2006): 481-486. https://doi.org/10.1109/FGR.2006.35.

[17] Yanan Qin, Tao Huang and Guanzheng Tang. "A Hierarchical Children's Dance Movement Pose Estimation Method Based on Sequence Multiscale Feature Fusion Representation." Advances in Multimedia (2022). https://doi.org/10.1155/2022/2445210.

[18] G. Qian, Feng Guo, T. Ingalls, L. Olson, J. James and T. Rikakis. "A gesture-driven multimodal interactive dance system." 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), 3 (2004): 1579-1582 Vol.3. https://doi.org/10.1109/ICME.2004.1394550.

[19] Abe Davis and Maneesh Agrawala. "Visual rhythm and beat." ACM Transactions on Graphics (TOG), 37 (2018): 1 - 11. https://doi.org/10.1145/3197517.3201371.

[20] Yi-Huang Su and Elvira Salazar-López. "Visual Timing of Structured Dance Movements Resembles Auditory Rhythm Perception." Neural Plasticity, 2016 (2016). https://doi.org/10.1155/2016/1678390.

[21] Mehdi Fallahpour and D. Megías. "High capacity audio watermarking using the high frequency band of the wavelet domain." Multimedia Tools and Applications, 52 (2011): 485-498. https://doi.org/10.1007/s11042-010-0495-1.

[22] Mehdi Fallahpour and D. Megías. "Secure logarithmic audio watermarking scheme based on the human auditory system." Multimedia Systems, 20 (2014): 155-164. https://doi.org/10.1007/s00530-013-0325-1.

[23] Xian-min Zhai. "Dance Movement Recognition Based on Feature Expression and Attribute Mining." Complex., 2021 (2021): 9935900:1-9935900:12. https://doi.org/10.1155/2021/9935900.

[24] Marc Gowing, Philip Kelly, N. O'Connor, Cyril Concolato, S. Essid, J. L. Feuvre, R. Tournemenne, E. Izquierdo, V. Kitanovski, Xinyu Lin and Qianni Zhang. "Enhanced visualisation of dance performance from automatically synchronised multimodal recordings." Proceedings of the 19th ACM international conference on Multimedia (2011). https://doi.org/10.1145/2072298.2072414.

[25] Nicole Harbonnier-Topin and J. Barbier. "“How seeing helps doing, and doing allows to see more”: the process of imitation in the dance class." Research in Dance Education, 13 (2012): 301 - 325. https://doi.org/10.1080/14647893.2012.677423.

[26] Huma Chaudhry, K. Tabia, Shafry Abdul Rahim and S. Benferhat. "Automatic annotation of traditional dance data using motion features." 2017 International Conference on Digital Arts, Media and Technology (ICDAMT) (2017): 254-258. https://doi.org/10.1109/ICDAMT.2017.7904972.

[27] Hyun-Tae Kim, Eungyeong Kim, Jong-Hyun Lee and C. Ahn. "A recommender system based on genetic algorithm for music data." 2010 2nd International Conference on Computer Engineering and Technology, 6 (2010): V6-414-V6-417. https://doi.org/10.1109/ICCET.2010.5486161.

[28] Hung-Chen Chen and Arbee L. P. Chen. "A Music Recommendation System Based on Music and User Grouping." Journal of Intelligent Information Systems, 24 (2005): 113-132. https://doi.org/10.1007/s10844-005-0319-3.

[29] Kunsu Kim, Dong-Hong Lee, T. Yoon and Jee-Hyong Lee. "A music recommendation system based on personal preference analysis." 2008

First International Conference on the Applications of Digital Information and Web Technologies (ICADIWT) (2008): 102-106. https://doi.org/10.1109/ICADIWT.2008.4664327.

[30] Longjun Huang, Li-pin Dai, Y. Wei and Minghe Huang. "A Personalized Recommendation System Based on Multi-agent." 2008 Second International Conference on Genetic and Evolutionary Computing (2008): 223-226.
https://doi.org/10.1109/WGEC.2008.45.

[31] Pradnya V. Kulkarni, S. Rai and Rohini S. Kale. "Recommender System in eLearning: A Survey." (2020): 119-126.
https://doi.org/10.1007/978-981-15-0790-8_13.

[32] Yazhou Yao, Jian Zhang, Fumin Shen, Dongxiang Zhang, Zhenmin Tang and Heng Tao Shen. "Towards Automatic Construction of Diverse, High-Quality Image Datasets." IEEE Transactions on Knowledge and Data Engineering, 32 (2017): 1199-1211.
https://doi.org/10.1109/TKDE.2019.2903036

[33] Shuhei Tsuchida, Satoru Fukayama and Masataka Goto. "Query-by-Dancing: A Dance Music Retrieval System Based on Body-Motion Similarity." (2018): 251-263.
https://doi.org/10.1007/978-3-030-05710-7_21.

[34] Alfdaniels Mabingo. "Intercultural Dance Education in the Era of Neo-State Nationalism." Journal of Dance Education, 19 (2019): 47 - 57.
https://doi.org/10.1080/15290824.2018.1434527.

[35] Sourya Dey, Kuan-Wen Huang, P. Beerel and K. Chugg. "Pre-Defined Sparse Neural Networks With Hardware Acceleration." IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 9 (2018): 332-345.
https://doi.org/10.1109/JETCAS.2019.2910864.

[36] R. Cai, Taihao Han, Wenyu Liao, Jie Huang, Dawang Li, Aditya Kumar and Hongyan Ma. "Prediction of surface chloride concentration of marine concrete using ensemble machine learning." Cement and Concrete Research, 136 (2020): 106164. https://doi.org/10.1016/j.cemconres.2020.106164.

[37] B. Pavlyshenko. "Using Stacking Approaches for Machine Learning Models." 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP) (2018): 255-258. https://doi.org/10.1109/DSMP.2018.8478522

[38] Meshia Cédric Oveneke, V. Enescu and H. Sahli. "Real-Time Dance Pattern Recognition Invariant to Anthropometric and Temporal Differences." (2012): 407-419.
https://doi.org/10.1007/978-3-642-33140-4_36.

[39] N. Magnenat-Thalmann, D. Protopsaltou and E. Kavakli. "Learning How to Dance Using a Web 3D Platform." (2007): 1-12. https://doi.org/10.1007/978-3-540-78139-4_1.

[40] Shuai Zhang. "An Intelligent and Fast Dance Action Recognition Model Using Two-Dimensional Convolution Network Method." Journal of Environmental and Public Health, 2022 (2022). https://doi.org/10.1155/2022/4713643.