Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/76/20240580

A comprehensive review of AI-based dance generation applications and their developmental prospects in virtual reality

Zheqing Zhang

Faculty of Electrical Engineering, Information Technology, Physics, TU Braunschweig, Braunschweig, Niedersachsen, Germany

kurtinas.ppq@gmail.com

Abstract. In the context of the rapid development of large AI models, multimodal fusion, and extensive application of VR, particularly in the era where technology profoundly impacts artistic expression, this review explores the innovative intersection of artificial intelligence (AI), virtual reality (VR), and dance choreography. It discusses the evolution of dance creation from traditional methods to advanced AI-driven techniques, emphasizing the role of AI in analyzing and synthesizing complex dance movements. The review highlights the application of multimodal learning in dance, focusing on how AI utilizes auditory and visual data to understand and create dance sequences. A significant part of the review is dedicated to the integration of VR in dance, examining its potential to enhance the choreographic process and user experience. The paper also addresses the challenges and future prospects in AI choreography, including the development of VR interfaces for dance enthusiasts and creators, akin to modern music production software. This venture reflects a promising direction for AI and VR in transforming dance creation and presentation, making it more accessible and interactive.

Keywords: Multimodal Learning, Human-Computer Interaction, Motion Capture, Interactive Design, Dance Choreography.

1. Introduction

In the contemporary artistic landscape, technological advancements in AI models, such as those seen in Magenta Studio, are rapidly transforming creative fields like music and painting. Advanced technologies like Differentiable Digital Signal Processing (DDSP) and Generative Adversarial Network Synthesis (GANSynth) are increasingly being explored, revolutionizing AI in music composition and representing the current new trend of combining classic signal processing technology with deep learning [1-2]. The integration of Ableton Live with Magenta Studio epitomizes this evolution, democratizing music composition and making it accessible to the general public. These advanced models are increasingly serving as invaluable tools for creators, augmenting their artistic capabilities.

Given the burgeoning development of multimodal artificial intelligence (AI), the immense potential of a sophisticated dance model is soon to be noticed. Currently, models like Chor-rnn utilize deep learning to create complex dance sequences, while applications such as GrooveNet and AI Choreographer that focus on real-time music-driven dance movement generation are gradually

DOI: 10.54254/2755-2721/76/20240580

becoming mainstream [3-4]. In light of these developments, we can anticipate the emergence of more user-friendly interactive interfaces in the future, which will have a revolutionary impact on dance creation for ordinary users. By simplifying the process of dance composition, these technologies not only ignite the public's passion for creation but also unleash their creative potential. This progress signifies the potential role of AI in dance artistry, offering opportunities for creative expression to a broader audience.

2. Technological foundations and advancements in dance movement recognition

2.1. From biomechanics to dance: the evolution of motion recognition

The journey of AI choreography research can be traced back to the study of visual perception of biological motion. In the early 20th century, Edgar Rubin conducted research on how people perceive shapes and patterns, focusing on the separation of figure-background in visual perception. He introduced the famous "Rubin Vase" or "Rubin's face-vase" illusion [5]. Subsequently, Albert Michotte pointed out that the visual system utilizes universal mechanisms to interpret observed dynamic events, emphasizing the ability of people to directly perceive causality without any reasoning or logical analysis, and to recognize specific biological motion patterns such as walking and running from abstract events [6]. These studies, from a psychological perspective, provided a theoretical foundation for identifying complex movements through the visual system. Later, Johansson developed a method by representing the movement of major joints in an active body with a few bright spots, thus studying motion patterns independently of shape. It was found that even a minimal number of 10-12 elements, through appropriate motion combinations, could evoke a strong visual impression of human movements such as walking, running, and dancing [7].

By the 21st century, in the pioneering stages of behavior recognition, Oliver tackled the challenge of limited training data by introducing and evaluating two state-based learning architectures—HiddenMarkov Models (HMM) and Coupled Hidden Markov Models (CHMM), aimed at modeling behaviors and interactions. Building on the foundation of time series modeling, CHMMs enrich the analysis by weaving interactions between multiple sequences, making them particularly adept at handling scenarios marked by sparse data [8]. To further advance this approach, Oliver and his team embarked on another innovative research project, leveraging a synthetic Alife-style training system. This system was designed to cultivate flexible prior models adept at recognizing human interactions, skillfully marrying Bayesian methods with existing knowledge and data evidence, thereby eliminating the need for extensive recalibration or additional training. Their groundbreaking work laid down a robust framework that empowered AI to effectively navigate small datasets and adapt to a wide range of complex behaviors—surpassing simple observations of passers-by in shopping malls or individuals walking past elevators.

Subsequently, Brand and Hertzmann's "Style Machines" marked a significant shift. They used motion capture sequences to replicate and learn various dance styles, allowing for the creation of new dance movements [9]. This progress highlights the transformative role of AI in dance choreography.

2.2. Opportunities and challenges in complex movement recognition

Development of motion capture starts with recognizing simple to complex human actions. While not directly centered on dance movements, it has provided invaluable insights and technological advancements.

Pavlovic introduced three algorithms for inference in Switching Linear Dynamic Systems (SLDS), treating them as dynamic Bayesian networks and demonstrating their superiority over traditional Hidden Markov Models in classification and prediction [10]. Ronald Poppe's 2010 survey addressed challenges in human action recognition of complex sports performance, and examined cost-effective image representation methods like global grid-based and space-time interest point techniques, enhancing human motion analysis [11]. This advancement is crucial for AI's understanding of complex dance movements, especially those with intricate motions and expressions.

DOI: 10.54254/2755-2721/76/20240580

As complex motion capture technologies mature, the volume and complexity of human motion data have increased significantly. Dance, being a highly structured spatiotemporal art form, requires precise coordination in time and space. In 2015, Holden introduced a novel technique using convolutional autoencoders to learn the manifold of human motion data. This method was applied to the extensive CMU human motion database, compressing complex human motion data into lower-dimensional representations and optimizing the efficiency of motion capture [12]. In 2016, Jain and others demonstrated the advantages of Structured Recurrent Neural Networks (S-RNN) over unstructured (conventional) RNNs and non-deep learning structured methods in various spatiotemporal problems. They transformed arbitrary spatiotemporal graphs into rich, scalable, and jointly trainable RNN hybrids [13]. These studies have significantly improved the quality of the data, constantly refining and optimizing human motion data and enabling prediction over extended periods of motion capture.

To process the broad spectrum of motion capture data, Butepage et al. developed a deep learning framework with an encoding-decoding network. They introduced an unsupervised learning scheme for the long-term prediction of human movement, proposing a Temporal Encoder (TE) aimed at capturing the temporal correlations in human motion data, rather than just static representations of human posture [14]. In their experiments, Butepage et al. demonstrated three distinct structures of Temporal Encoders: the S-TE model utilizes a symmetric structure for simultaneous encoding and decoding of motion data; the C-TE model improves predictive ability by accounting for motion data across different time scales; and the H-TE model achieves more precise capture and prediction of complex human actions by directly incorporating the body's hierarchical structure. Such methods, by offering multiple perspectives on action prediction and encoding, provide a solid technical foundation for the prediction and generation of dance movements, which require high coherence and temporal correlation.

2.3. Multimodal approaches and emotion analysis

Dance, as a multifaceted art form, encompasses music, movement, lighting, and emotional expression. Therefore, the integration of multimodal fusion in AI for dance recognition can enhance the processing and interpretation of dance data. Atrey et al. explain that multimodal fusion combines multiple sensory inputs, like audio and visual data, for a more comprehensive understanding, which is essential in dance, where auditory and visual elements are closely intertwined [15]. Baltrusaitis et al. note the challenge of synchronizing different mediaˇ types, such as audio and video, in multimodal fusion [16]. Moreover, assessing the confidence levels of different modalities is crucial in dance recognition. Sometimes audio cues may be more reliable than visual ones, necessitating a weighted fusion approach [15]. The decision-making in multimodal fusion affects dance recognition's accuracy and efficacy. Lastly, determining which modalities to fuse, as they can provide complementary or contradictory insights, is key for effective fusion in dance recognition.

Back to 2004, Camurri's research developed a set of algorithms and software modules focused on real-time analysis of expressive gestures in full-body human motion [17]. Subsequently, various methods for analyzing the expressiveness of gestures were proposed, highlighting the role of non-propositional movement traits (such as amplitude, speed, and fluidity) in emotional inference [18]. For a more detailed interaction case, Sanghvi proposed a method based on automatic visual extraction of expressive posture features from side-view video captures, utilizing computer vision to extract features such as body lean angle, slouch factor, quantity of motion, and contraction index. These features are further transformed into metafeatures for the training of recognition models, facilitating the automatic prediction of subjects' engagement [19]. The study specifically selected children interacting with an iCat robot in a chess game as subjects, providing a rich dataset for training and validation due to the pronounced and diverse expressive postures and body movements of children compared to adults. Later, Kleinsmith and Bianchi Berthouze identified different roles of body posture and movement information in non-verbal communication for emotion perception. They explored the relationships between emotional states and high-level descriptions of body movements or postures, as well as how to refine emotional body expressions into lower-level descriptions. Their research provided a theoretical basis for developing AI systems capable of recognizing and mimicking emotional expressions in dance [20].

DOI: 10.54254/2755-2721/76/20240580

3. Unlocking potential in AI dance generation

3.1. Comprehensive overview of AI synthesis

Artificial Intelligence (AI) is making significant strides in synthesizing various forms of content, including textual, auditory, and visual creations. Without the need for explicit supervision on specific tasks, GPT has been capable of autonomously performing a range of tasks, from text understanding, translation, and summarization to answering questions, flexibly achieving zero-shot task transfer across most training scenarios [21]. WaveNet, through the integration of dilated and causal convolutions, models and generates raw audio waveforms directly without relying on preprocessing or manual feature extraction. This approach achieves a high degree of naturalness in synthesizing and transitioning between various audio samples, such as naturally switching between different languages in the simulation of human speech or stitching together music segments, continually improving the quality of generated audio [22-23]. The text-guided diffusion model GLIDE employs a classifier-free guidance strategy, enabling the generation of images by DALL-E that closely resemble real photographs and are more favored by humans in evaluations of realism [24]. Furthermore, the direct generation of video files from textual descriptions is also making preliminary progress. These advancements seem to indicate that ordinary users are only one step away from accessing a user-friendly AI dance generation interface in their daily life.

3.2. Music-driven dance generation

In the context of AI-driven dance generation for dance application scenarios, text inputs often contain more abstract information, making the understanding of user intent a significant challenge today. On the other hand, audio-driven approaches serve as a more effective starting point for simple dance generation due to the higher correspondence between musical and dance features, offering a richer model for training. Ofli first proposed an innovative framework employing four models: a music measurement model, an exchangeable dance character model, a dance character transformation model, and a dance character model to achieve a many-to-many statistical mapping [25]. Based on this, researchers defined a discrete HMM and synthesized different dance sequences using a modified Viterbi algorithm. Then, the motion parameters of the synthesized dance were calculated using the dance figure models, and these parameters were animated synchronously with the musical audio using a 3D character model.

The matching of music and dance has always been a focus of research on music-to-dance generation. In Ofli's research, AI learns exchangeable groups of dance figures, comparing and capturing the intrinsic dependencies within sequences of dance figures, allowing for acceptable variations within the training model. This enhances the richness of the generated movements while ensuring continuity of action and structural consistency [25]. To bring the generated dances closer to reality, the MDOT-Net (Music-to-Dance with Optimal Transport Network) framework incorporates the optimal transport distance and the Gromov-Wasserstein distance to measure the similarity between distributions across different domains, such as the music space and the dance posture manifold. This effectively solves the cross-domain generation problem, ensuring that the dance sequences align well with the given music input in style and rhythmically match the expression [26]. Subsequently, HY Au's team focused on the aesthetic style information between motion and music, exploring a dynamic graph-based data-driven learning strategy. Through training, a music style embedding module was developed, concentrating on selections sensitive to changes in music style. The strategy dynamically adjusts the speed of selected dance motion segments according to changes in the music, incorporating the completeness of actions, the naturalness of transitions, and the appropriateness of style into the choreography process . This framework flexibly responds to changes in music style, adjusting the selection and combination strategy of nodes composed of dance segments, to generate dances that match the music style [27].

3.3. Current progress and challenges in AI dance generation

In recent years, the Chor-RNN (Choreography Recurrent Neural Network) has been a focal model for dance generation. By adapting variations of RNNs (Recurrent Neural Networks), such as LSTM (Long

DOI: 10.54254/2755-2721/76/20240580

Short-Term Memory) and GRU (Gated Recurrent Unit), it excels in handling time-series data, including music, language, and motion sequences. Compared to traditional rule-based dance synthesis systems, such as those using Markov models for music composition , this approach exhibits greater flexibility and creativity. Chor-RNN not only generates new dance sequences but also maintains and extends the dance style and expressiveness, surpassing previous systems in diversity and creativity. Additionally, by integrating Mixture Density Networks (MDN), the output's vitality and variety are further enhanced, overcoming potential stagnation in dance fragment output [28].

Despite significant advancements in dance synthesis, an increasing number of applications for dance synthesis are emerging. Training dance models remains unstable and complex, often involving abstract, high-dimensional information. Mature dance generation systems like GrooveNet, AI Choreographer, and the AIST++ Dataset still face the challenge of limited training data. The GrooveNet project utilizes FCRBM and LSTM-RNN, along with different methods for describing audio information (feature extraction and feature learning), hoping the model can learn independently from audio data and generate basic dance moves, thus generating more training samples. However, it struggles to generalize beyond the songs in the training data, showing a tendency towards overfitting [3].

Choreography Optimization is predicted be a more advanced trend through the application of the optimal transport framework, aiming for generality and composability of dance content. However, the generalization capability of dance training models remains a significant challenge. Dance, an activity that combines complex spatial and temporal sequences, involving intricate human movements, temporal relationships between movements, rhythm, and dance style, among other multidimensional information. Progress in this field may only be achieved by capturing more synchronized dance and music data, deepening the understanding of various information interactions in dance scenarios, such as lighting, the environment, and integrating a broader range of multimodal information, to make future breakthroughs.

4. Virtual reality and AI dance: advancements and future prospects in interactive tools

Popular interactive dance tools like "Just Dance" utilize handheld controllers for operation, capturing users' movements during play. These tools leverage a powerful interface, ambiance leadership, and controller feedback, offering most users a sense of strong interactivity and engagement [29]. However, the software relies on capturing professional dancers' movements directly, leading to high costs and complex processes. Resources require manual replenishment and maintenance, and users need to invest in memberships for long-term use. Additionally, simulated dance learning through controllers does not accurately reflect the nuances and skills of real dancing.

"Dancing in the Streets" (DITS), one of the earliest wearable VR products designed specifically for dance exercise, aims to replace traditional dance mats. Its immersive VR experience and the application's gamification and social aspects significantly motivate players to continue playing [30]. With VR technology's rapid development, advanced time-warping techniques optimally match image generation with user perception through algorithms, making VR experiences more immersive and natural. Games like "Beat Saber" that encourage full-body movement represent the pinnacle in user-application interaction. These developments demonstrate the potential of dance interaction in VR and the trend towards choreography software dominated by VR interfaces. As VR platforms become more widespread and VR communities more refined, the way people create dance could be fundamentally transformed.

5. Conclusion

The emergence of AI choreography tools, such as AI Choreographer and the AIST++ Dataset, signifies a growing interest in the AI choreography market. After prolonged development, from the visual perception of biological motion to understanding complex human behaviors, AI choreography is finally on the verge of having its own workstation, just like various DAWs(Digital Audio Workstation) for music production. Despite numerous limitations in processing complex information at present, the development of large AI models, the continuous expansion of training data, and the increasing interactivity and usability of VR are laying a solid foundation for its progress.

DOI: 10.54254/2755-2721/76/20240580

Future research should focus on developing more efficient and versatile artificial intelligence systems that integrate multimodal data (such as sound, motion, light, and emotions) to explore more multimodal information that can be processed, aiming to enhance the quality of dance choreography and the innovativeness of the output content. AI-assisted choreography software on VR platforms is poised to elevate dance creation to new heights, offering new tools for dance artists and broadening possibilities for dance education and entertainment.

References

[1] Engel, Jesse, et al. ”DDSP: Differentiable digital signal processing.” arXiv preprint arXiv:2001.04643 (2020).
[2] Engel, Jesse, et al. ”Gansynth: Adversarial neural audio synthesis.” arXiv preprint arXiv:1902.08710 (2019).
[3] Crnkovic-Friis, Luka, and Louise Crnkovic-Friis. ”Generative choreography using deep learning.” arXiv preprint arXiv:1605.06921 (2016).
[4] Alemi, Omid, Jules Franc¸oise, and Philippe Pasquier. ”GrooveNet: Realtime music-driven dance movement generation using artificial neural networks.” networks 8.17 (2017): 26.
[5] Rubin, Edgar. Visuell wahrgenommene figuren: Studien in psychologischer analyse. Vol. 1. Gyldendalske boghandel, 1921.
[6] Mays, W. ”The Perception of Causality.” (1964): 254-259.
[7] Johansson, Gunnar. ”Visual perception of biological motion and a model for its analysis.” Perception psychophysics 14 (1973): 201-211.
[8] Oliver, Nuria M., Barbara Rosario, and Alex P. Pentland. ”A Bayesian computer vision system for modeling human interactions.” IEEE transactions on pattern analysis and machine intelligence 22.8 (2000): 831-843.
[9] Brand, Matthew, and Aaron Hertzmann. ”Style machines.” Proceedings of the 27th annual conference on Computer graphics and interactive techniques. 2000.
[10] Pavlovic, Vladimir, James M. Rehg, and John MacCormick. ”Learning switching linear models of human motion.” Advances in neural information processing systems 13 (2000).
[11] Poppe, Ronald. ”A survey on vision-based human action recognition.” Image and vision computing 28.6 (2010): 976-990.
[12] Holden, Daniel, et al. ”Learning motion manifolds with convolutional autoencoders.” SIGGRAPH Asia 2015 technical briefs. 2015. 1-4.
[13] Jain, Ashesh, et al. ”Structural-rnn: Deep learning on spatio-temporal graphs.” Proceedings of the ieee conference on computer vision and pattern recognition. 2016.
[14] Butepage, Judith, et al. ”Deep representation learning for human motion prediction and classification.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[15] Atrey, Pradeep K., et al. ”Multimodal fusion for multimedia analysis: a survey.” Multimedia systems 16 (2010): 345-379.
[16] Baltrusaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency.ˇ ”Multimodal machine learning: A survey and taxonomy.” IEEE transactions on pattern analysis and machine intelligence 41.2 (2018): 423-443.
[17] Camurri, Antonio, Barbara Mazzarino, and Gualtiero Volpe. ”Analysis of expressive gesture: The eyesweb expressive gesture processing library.” Gesture-Based Communication in Human-Computer Interaction: 5th International Gesture Workshop, GW 2003, Genova, Italy, April 15-17, 2003, Selected Revised Papers 5. Springer Berlin Heidelberg, 2004.
[18] Castellano, Ginevra, Santiago D. Villalba, and Antonio Camurri. ”Recognising human emotions from body movement and gesture dynamics.” International conference on affective computing and intelligent interaction. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

DOI: 10.54254/2755-2721/76/20240580

[19] Sanghvi, Jyotirmay, et al. ”Automatic analysis of affective postures and body motion to detect engagement with a game companion.” Proceedings of the 6th international conference on Human-robot interaction. 2011.
[20] Kleinsmith, Andrea, and Nadia Bianchi-Berthouze. ”Affective body expression perception and recognition: A survey.” IEEE Transactions on Affective Computing 4.1 (2012): 15-33.
[21] Radford, Alec, et al. ”Language models are unsupervised multitask learners.” OpenAI blog 1.8 (2019): 9.
[22] Van Den Oord, Aaron, et al. ”Wavenet: A generative model for raw audio.” arXiv preprint arXiv:1609.03499 12 (2016).
[23] Gibiansky, Andrew, et al. ”Deep voice 2: Multi-speaker neural text-tospeech.” Advances in neural information processing systems 30 (2017).
[24] Nichol, Alex, et al. ”Glide: Towards photorealistic image generation and editing with text-guided diffusion models.” arXiv preprint arXiv:2112.10741 (2021).
[25] Ofli, Ferda, et al. ”Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis.” IEEE Transactions on Multimedia 14.3 (2011): 747-759.
[26] Wu, Shuang, Shijian Lu, and Li Cheng. ”Music-to-Dance Generation with Optimal Transport.” arXiv preprint arXiv:2112.01806 (2021).
[27] Au, Ho Yin, et al. ”Choreograph: Music-conditioned automatic dance choreography over a style and tempo consistent dynamic graph.” Proceedings of the 30th ACM International Conference on Multimedia. 2022.
[28] Crnkovic-Friis, Luka, and Louise Crnkovic-Friis. ”Generative choreography using deep learning.” arXiv preprint arXiv:1605.06921 (2016).
[29] Lin, Jih-Hsuan. ”“Just Dance”: the effects of exergame feedback and controller use on physical activity and psychological outcomes.” Games for health journal 4.3 (2015): 183-189.
[30] Clawson, James, Nirmal Patel, and Thad Starner. ”Dancing in the Streets: The design and evaluation of a wearable health game.” International Symposium on Wearable Computers (ISWC) 2010. IEEE, 2010.