DanceVision Proposal

Anna Guo Jenna Jacob Katie Wang Natalie Leal Blanco
UT Austin: Computer Vision Course
annayuab@icloud.com

Abstract

Dance is a universal form of expression and entertainment, yet mastering its intricacies can be daunting for beginners. To democratize dance learning and make it more accessible to enthusiasts, we propose a novel approach utilizing computer vision techniques. By breaking down dance videos into vectorized and text-based representations with counts, the complexities of dance routines are simplified, enabling easier comprehension and replication by average fans.

1. Introduction

In recent years, dance has surged in popularity, whether through TikTok trends in younger generations, the growing fanbase for elaborate K-pop dance videos, or the performances in reality competition shows (which themselves cover wide genres from ballet and contemporary to ballroom dances). As a result of this growing interest, many eager dance novices have attempted to self-teach from the abundant dance videos online, available through a variety of platforms. However, such individuals often struggle with breaking down movements due to the speed or intricacy with which they were executed.

We aim to use Computer Vision to partition dance movements into more easily digestible chunks by creating a model that can learn (vectorize) dance choreography from a given dance video, and output it in a more human-readable form, such as by describing multiple steps in a motion and adding counts or timings for context. By developing a tool that provides detailed descriptions of human positions and gestures alongside counts or musical beats, learning dance choreographies could become much more accessible and approachable for a broader audience.

1.1. Existing Attempts

Several existing attempts have been made to tackle the challenge of learning dance from online videos, each with its own approach and limitations. One such approach involves platforms that generate reskinned versions of dance performances, offering visual references for users. These platforms overlay graphics or animations onto the original video, highlighting key movements or positions to provide guidance. While these reskinned versions serve as visual aids, they typically do not offer active instruction or detailed breakdowns of the dance steps. As a result, users may struggle to fully grasp the nuances and intricacies of the movements, limiting the effectiveness of this approach as a learning tool.

Another example is a DancePal Hackathon project, which adopts a more technical approach by comparing an ideal video with the end user's video skeletons[5]. This method involves analyzing the skeletal poses extracted from both videos and identifying discrepancies between them. While this approach has the potential to provide personalized feedback to users based on their own performances, it comes with several challenges. Firstly, it requires an ideal video with specific lighting, background, and perspective conditions to serve as a reference, which may not always be feasible to obtain. Additionally, accurately mapping movements onto predefined definitions can be difficult, especially for complex dance sequences with varying styles and interpretations. As a result, the effectiveness of this approach may be limited by the availability of suitable reference videos and the accuracy of the mapping process.

Unlike reskinned versions of dance performances, our proposed method provides detailed instructional breakdowns of dance movements, guiding learners through each step with clarity and precision. This active instruction facilitates better understanding and retention of choreography. Additionally, our method could be expanded to offer personalized feedback based on the analysis of the learner's own video and that of an original dance video, ideally without the requirement for perfect lighting or camera angle conditions.

1.2. Useful Techniques

Several computer vision techniques offer promising avenues for addressing the challenge of learning dance from online videos by providing deeper insights into the dynamics and nuances of dance movements. One such technique involves exploring 3D models instead of traditional 2D skeletonization methods to create more comprehensive representations of dance movements. Unlike 2D skeletonization, which captures only the spatial coordinates of key body joints in a single plane, 3D modeling techniques can capture the full spatial extent of movements, including depth information[1]. This allows for a more accurate and detailed representation of the dancer's body and movements in three-dimensional space, which is particularly valuable for capturing complex and multi-dimensional dance choreographies[12].

In addition to 3D modeling, other computer vision techniques such as action recognition, pose estimation, trajectory analysis, and video segmentation and tracking can provide valuable insights into the dynamics of dance movements. Action recognition algorithms can automatically identify and classify different dance gestures and sequences, allowing for automated analysis of choreographic patterns and styles. Pose estimation techniques can accurately estimate the skeletal poses of dancers from video frames, providing a structural representation of their movements that can be used for further analysis and interpretation.

Trajectory analysis techniques can track the trajectories of specific body parts or movements over time, allowing for the analysis of movement patterns, velocities, and accelerations[2]. This can provide valuable insights into the rhythmic and dynamic aspects of dance performances, helping learners to better understand the timing and pacing of movements. Video segmentation and tracking methods can segment dance videos into individual sequences or movements, allowing for more focused analysis and comparison of specific choreographic elements.

By leveraging these advanced computer vision techniques, learners can gain a deeper understanding of the dynamics and nuances of dance movements, facilitating more effective learning experiences. These techniques can help learners to break down complex choreographies into smaller, more manageable components, identify areas for improvement, and refine their technique with greater precision and accuracy. Overall, the integration of computer vision into dance learning has the potential to revolutionize the way dance is taught and learned, making it more accessible, engaging, and effective for dancers of all levels.

1.3. Challenges

Several challenges arise from attempting this project. One is the quality of input data, which encompasses various factors such as lighting, angles, occlusions, frame rates, and blurring. These factors can significantly impact the accuracy of movement analysis, as they can obscure or distort key visual information needed for understanding dance movements. Addressing these challenges may require the development of robust computer vision algorithms capable of handling noisy or low-quality video data, as well as techniques for preprocessing and enhancing the visual information to improve analysis accuracy.

Moreover, handling interactions between multiple dancers in an ideal video presents additional complexities in accurately capturing and representing dance movements. In many dance performances, especially group or partner dances, dancers may interact with each other in complex ways, such as lifts, partner work, or coordinated movements. Capturing these interactions and accurately representing them in the analysis poses challenges for computer vision algorithms, as they must be able to differentiate between individual dancers and track their movements independently while also accounting for their interactions with others. Addressing this challenge may require the development of advanced tracking and segmentation techniques capable of accurately identifying and tracking multiple dancers in a crowded or dynamic environment.

Overall, addressing these challenges requires a multidisciplinary approach that combines expertise in computer vision, dance analysis, and human-computer interaction. By developing robust algorithms, incorporating domain-specific knowledge, and leveraging advanced techniques for handling complex visual data, it is possible to create a tool that effectively supports learning dance from online videos using computer vision. However, continued research and development efforts are needed to overcome these challenges and realize the full potential of computer vision technology in the field of dance education.

1.4. Technical Gap

The primary aim of this project is to reduce the technical gap between online dance learning resources and effective dance instruction. By enabling the correspondence between individuals with different body proportions captured from various angles, this tool seeks to enhance the accessibility and inclusivity of dance education. This interdisciplinary endeavor involves collaboration between experts in dance, biomechanics, and computer vision, with computer vision providing quantitative analysis capabilities to enhance the understanding and teaching of dance movements.

In conclusion, developing a tool for learning dance from online videos using computer vision holds great potential for revolutionizing dance education and making it more accessible to a wider audience. By leveraging advanced computer vision techniques and interdisciplinary collaboration, this project aims to bridge the gap between online dance resources and effective dance instruction, ultimately empowering individuals to learn and appreciate dance in new and innovative ways.

2. Pose Estimation Models

The first step in developing this project was to define the parameters and requirements in order to choose models and algorithms that satisfy our expectations. The first model that will be implemented on our project is a Pose Estimation Model. The requirements we defined were: limit GPU usage, multiple persons detection, and output in the form of JSON files, in order to be transferable and readible for the subsequent sections of our project. We also set a high priority on accuracy, due to the requirements for the other portions of our project[10]. The following section will highlight our research, testing, and challenges, as well as the factors that influenced our decision-making process.

2.1. Model Choices

In order for our project to work as intended, we needed to make sure we started by using the optimal tools for data collection. Our project was in need of a pose estimation model, and some options that we considered were OpenPose, MediaPipe, and MoveNet.

1. MediaPipe: MediaPipe is a Framework with many applications that have been widely used by Google for numerous products and services[11]. For this project specifically, MediaPipe would aid specifically in pose detection and video processing using OpenCV. Unfortunately, this is limited to single-person analysis, and we wanted to account for the possibility of expanding our project to process videos with multiple people.

2. MoveNet: MoveNet is another pose estimation model that we considered. Developed by Google, it utilizes deep learning techniques, and is an improved version of the previous generation PoseNet model (released in 2017)[9]. Though this processed our videos quickly, it did not have the desired accuracy for the video inputs (key points tended to be in the general region of the desired body parts but didnt always overlap fully, which led to doubts on its ability to capture the fine granularity we wanted for detailed dance movements).

3. OpenPose: OpenPose is yet another software for pose estimation, which has proved to work well when evaluating multiple persons, even for more crowded scenes. It uses a bottom-up approach to produce higher quality results, though at a higher computational cost compared to MediaPipe and MoveNet[6]. Additionally, it outputs JSON files, one per video frame, containing data for keypoints (e.g., left knee, right knee, etc.) that were easier to process.

Because of its high accuracy and usable output, OpenPose was selected for this project over MediaPipe and MoveNet.

2.2. Challenges with OpenPose

Initially, OpenPose appeared to require more significant computational resources than were readily available to us (RAM and GPU), but we were able to run the model successfully after closing other running programs on our computers and using the CPU-version of OpenPose.

As we became more familiar with OpenPose, we began testing the effects of different parameters. For instance, we learned that lowering the resolution of a video resulted in a significantly faster processing time. However, this also caused decreased accuracy in the pose estimation. Thus, after several rounds of testing, we decided against reducing the resolution in our input videos since our project performance doesnt depend on low latency, but does benefit from higher accuracy.

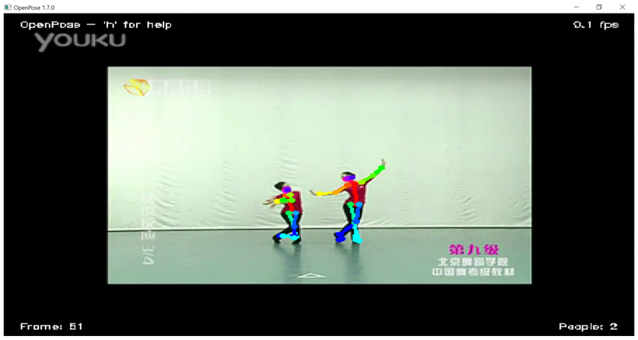

Figure 1. OpenPose in action: pose detection on a frame from an input video of two people dancing. OpenPose accurately detects the key points on both people despite their slightly unorthodox body positions.

3. Algorithm to Prepare OpenPose outputs for Feature Extraction

For a given video, OpenPose outputs a directory of json files, with one for each frame. Before passing these outputs to a second machine learning model, we want to write an algorithm that maps similar movements to the same values, regardless of camera angle or proximity of the person to the camera. (Due to the nature of OpenPose outputting x, y values for keypoints, we have already circumvented the issue of changes in illumination.) Our algorithm aims to:

1. Remove absolute position by calculating the change in positions of key points between consecutive frames.

2. Normalize position gradient using the maximum distance (in the 32 most recent frames) from right shoulder to right hip.

3. Accommodate consistent elliptical motion by calculating angles at body joints.

We choose to transition away from absolute positions to position gradients due to inspiration from SIFTs usage of illumination gradients. Absolute positions could cause our model to unintentionally train on a dancers location in the camera view rather than their movements, whereas position gradient should make our model invariant to the dancers x, y-positions.

Similarly, by attempting to normalize the position gradients to the length of the dancers body part, we hope to make our model invariant to the dancers height and their distance from the camera (people in the foreground tend to appear taller than people in the background). We choose the right shoulder and right hip specifically because we believe the distance in between to represent a relatively rigid part of the body with a length correlated to a persons height. Furthermore, we choose the longest length found to account for dance movements that may cause the shoulder and hip to be closer together than usual, while limiting the range to 2 seconds (32 frames) so that were still sensitive to global changes (changes throughout the video) in z position (the distance between the dancer and camera center).

Finally, we want to pass in angles calculated from the keypoints. Firstly, this will likely aid the model in considering elliptical motions, which are prevalent in dances, as a continuous motion, as opposed to a single elliptical motion being calculated as multiple straight-line movements. Additionally, this encodes information about body parts relative to each other. For instance, moving ones right hand downwards on the left side of ones body could be differentiated from moving ones right hand downwards on the right side of ones body through the angle of the right elbow relative to the right shoulder.

4. Model

After the initial preprocessing, our project will move on to using 3 additional machine learning models to extract feature points, create clusters, and classify dance genres. This will be done through creating a customized version of the SIFT algorithm, using DBSCAN for clustering and an undetermined machine learning model to label dance movements and the dance genre.

For feature extraction, we will implement a model that works similarly to SIFT, but is modified to work for videos instead of just images. The model will look at the data generated by the previously mentioned algorithm and extract the important dance moves from it. Similarly to how SIFT uses different sized kernels to make the model scale-invariant, we will make our model tempo-invariant by analyzing different frame lengths at a time to find the most accurate feature moves. We believe that a single dance move is categorized by a constant motion, so we can differentiate between two different dance moves by looking at when the dancer changes direction (or in the case of elliptical movements, changes angle gradient), similar to how the Harris Corner Detector works in images. Then, this data will be sent to a clustering algorithm.

To categorize the dance moves into clusters, we will utilize Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The model checks the neighborhood around the given point, and if it contains enough points (the minimum number of points being one of the hyperparameters of the function), then a cluster is started. Otherwise, the point is labeled as noise. In this way, the model can find clusters more accurately than other models while being knowledgeable about noise, which allows it to find arbitrarily shaped clusters. With this, we will be able to put the frames into clusters based on similar interest points. After this, the team will give labels to each cluster, so that this data can be used in the next step, which is to find the genre that the input dance belongs to.

The final step in the model process is to classify the genre of the dance. The dance move clusters generated by the DBSCAN will be analyzed against training data, and the output will be matching the given data with one of the dance genres from the training set. This will be done using an out-of-the-box machine learning model, whose specifics will be determined later. This model will most likely use supervised learning, since it uses the data we supply in the training phase as a baseline for categorizing new data.

5. Applications: Physical Therapy

Although this project originally had a focus on dance, we believe that it can be applied more widely in medical care, specifically in physical therapy. Since our project is able to read video inputs and detect the poses in videos, it can do the same for physical therapy patients. This would greatly serve patients in their Physical Therapy journeys, whether healing from a sports injury, recovering from a surgical procedure, or relieving pain from a specific condition. Patients undergoing this type of treatment will typically receive a plan outlining a set of exercises and stretches to be completed that aid in their process to restoration.

With the scope of our project, it can similarly analyze the data from these exercises and stretches being done, then evaluate the patients progress over time. This could help motivate the patient in their recovery and provide further insight for both the patient and their physical therapist as to how their recovery journey is progressing, also allowing for a better understanding of how ones plan might be adjusted accordingly. We aim for our project to have a level of flexibility that will allow for this to be easily adapted[4].

References

[1] A.Mumuni and F. Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Elsevier*, 16(1), 2022. 2

[2] B.B. Armor, J. Su, and A. Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):1–13, 2016. 2

[3] F. Baradel. *Structured deep learning for video analysis*. PhD thesis, 2004.

[4] S. Bernstein. What is physical therapy? *WebMD*, 2023. 4

[5] V. Buwaneka, A. Samadder, and K. Sharma. Dancevideo, 2023. 1

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3

[7] C. Chan, S. Ginosar, T. Zhou, and A. Efros. Everybody dance now, 2019.

[8] Y. Jafarian and H. Park. Learning high fidelity depths of dressed humans by watching social media dance videos, 2021.

[9] K. LeViet. Pose estimation and classification on edge devices with movenet and tensorflow lite, 2021. 3

[10] Maureentkt. Selecting your real-time pose estimation models. *Medium*, 2021. 3

[11] MediaPipe. Pose landmark detection guide, 2023. 3

[12] OpenCV. Open source computer vision library, 2024. 2

[13] Y. Pang and Y. Niu. Dance video motion recognition based on computer vision and image processing. *Applied Artificial Intelligence*, 37(1), 2023.

[14] M. Yang and Z. He. Dance action recognition model using deep learning network in streaming media environment. *PMC*, 2022.