CVIA - Assignment Report
Koi-un

Doruk Efe Kanber, Eren Karsavuranoglu, Tulga Kagan Temel,
3163051, 3164647, 3168178,

May 24, 2025

1 Introduction

This project addresses the challenge of *automated comparison of human dance movements* against a reference performance. As interactive applications like dance games and virtual fitness coaching become more prevalent, the need for robust methods to evaluate and provide feedback on human motion grows. The primary objective is to develop and analyze a system that can quantify the similarity between two dance performances by leveraging modern computer vision techniques.

2 Problem Formulation

The core problem is to define and measure "similarity" in a way that aligns with human perception of performance quality. Dance is highly nuanced, involving not just static poses but also rhythm, fluidity, and coordination. The importance of solving this problem lies in creating more engaging and effective interactive experiences. An accurate similarity score can provide valuable real-time feedback to a user, helping them improve their form and timing.

3 Main Contribution

The main contribution of this work is the development and comparative analysis of a system employing three distinct similarity metrics for dance evaluation. We implement and test Windowed Cosine Similarity (with magnitude scaling), a Mean Absolute Error (MAE), and a Correlation based score on a controlled dataset. Our analysis provides insights into the relative discriminative power and specific behaviors of these metrics.

4 Background

Our work builds upon established techniques in computer vision and time-series analysis. For pose estimation, we utilize Google's MediaPipe Pose framework [1], a state-of-the-art solution for real-time human pose tracking. MediaPipe provides 33 body keypoints, from which we selected 6 that are most relevant to dance choreography for joint angle calculation. We explore metrics like cosine similarity, MAE, and correlation, and adapt them for the specifics of joint angle time-series data.

5 Methodology

Our methodology follows a pipeline structure:

1. **Video Input:** Two videos (an avatar reference and a user performance) are provided.

2. **Pose Estimation:** Frames are extracted at a consistent rate. MediaPipePose [1] is used to detect 33 pose keypoints for each frame. Different parameters were used for extracting the keypoints of the avatar and the human.

3. **Feature Extraction:** Keypoints are normalized relative to hip center and shoulder width. Time series of key joint angles are then calculated.

4. **Similarity Computation:** Three metrics are used to compare the joint angle time series between the avatar and the user for overall performance assessment:

   - **Windowed Cosine Similarity (CosineSim):** Computes cosine similarity over sliding windows of the angle sequences. This is scaled by the ratio of vector magnitudes within the window to penalize differences in movement intensity. The final score is the average of these windowed, scaled similarities, expressed as a percentage.

- **MAE-based Similarity:** For sequences aligned by truncation, $S_{MAE} = \max(0, 1 - \text{MAE}/\Delta\theta_{max}) \times 100\%$, where $\Delta\theta_{max}$ is 180 degrees.
- **Correlation-based Similarity:** Uses the Pearson coefficient ρ between truncated sequences, $S_{Corr} = ((\rho + 1)/2) \times 100\%$.

An overall average score is computed for each metric. Additionally, for real-time visual feedback in the comparison video, dynamic MAE-based similarity scores are calculated over a sliding window.

6 Datasets

The data used for our experiments consists of three video recordings processed at a synchronized frame rate: an **Avatar Reference** video, a **Good Human Imitation**, and a **Bad Human Imitation**. This controlled set allows for testing the discriminative power of our similarity metrics.

7 Experiments and Results

Experiments were conducted comparing the avatar's performance against the "Good Imitation" and the "Bad Imitation." The objective was to assess if the implemented metrics could differentiate performance qualities. The results are summarized in Table 1 and Table 2.

The experimental results align with qualitative human assessment. Our main finding is that all three similarity metrics (Windowed Cosine Similarity, MAE-based Similarity, and Correlation based Similarity) successfully differentiated between the 'Good' and 'Bad' imitations, with the Good Imitation consistently achieving higher average scores.

Specifically, the "Good Imitation" outperformed the "Bad Imitation" with an average Cosine Similarity score of **86.4% vs 82.2%**, an average MAE-based similarity of **81.6% vs 80.1%**, and an average Correlation-based similarity of **63.8% vs 57.4%**. This validates the general effectiveness of the overall pipeline and the chosen features for capturing noticeable differences in performance quality. The dynamic scores displayed in the output video are based on windowed MAE. Visualizations are in Appendix Figures A.1 and A.2.

8 Discussion

The corrected results confirm that our methodology can effectively quantify and differentiate dance similarity. An analysis of the metrics reveals their varying sensitivities:

- **Relative Discriminative Power:** Correlation-based Similarity showed the largest margin (6.4 points) between the "Good" and "Bad" imitations. Windowed Cosine Similarity also provided good separation (4.2 points), while MAE-based Similarity had the smallest margin (1.5 points). This suggests that for our specific test data, the *pattern* of movement (captured by Correlation) and the combination of *local shape and intensity* (captured by CosineSim) were stronger differentiators of quality than average absolute angular error alone.
- **Metric Characteristics:**
  - **Cosine Similarity (Windowed with Magnitude Scaling):** This metric balances local shape similarity with intensity. Its effectiveness suggests that the "Bad Imitation" likely failed in both consistently matching local movement patterns and executing them with appropriate amplitude relative to the avatar.
  - **MAE-based Similarity:** While it did differentiate, its smaller margin might indicate that the "Bad Imitation" did not always exhibit drastically larger average angular errors across all frames compared to the "Good Imitation" (which might have had some larger, isolated deviations).
  - **Correlation-based Similarity:** Its strong discriminative performance implies that the "Bad Imitation" failed to replicate the temporal pattern of angle changes, even if other aspects were less consistently poor.

Table 1: Similarity Scores: Avatar vs Good Imitation (New Results).

Joint	**CosineSim (%)**	**MAE Sim. (%)**	**Corr. Sim. (%)**
Left Elbow	80.2	72.2	63.9
Left Hip	89.9	87.1	64.6
Left Knee	89.6	86.3	70.6
Right Elbow	79.7	71.3	60.3
Right Hip	89.5	86.8	63.1
Right Knee	89.6	85.8	60.3
Average	**86.4**	**81.6**	**63.8**

Table 2: Similarity Scores: Avatar vs Bad Imitation (New Results).

Joint	**CosineSim (%)**	**MAE Sim. (%)**	**Corr. Sim. (%)**
Left Elbow	72.3	68.5	50.3
Left Hip	88.1	86.7	61.1
Left Knee	86.2	84.6	59.1
Right Elbow	72.5	69.0	51.6
Right Hip	87.0	86.0	61.8
Right Knee	87.3	85.8	60.4
Average	**82.2**	**80.1**	**57.4**

- **Analysis of Score Magnitudes:** Even the ”Bad Imitation” received fairly high percentage scores from CosineSim and MAE-based metrics. This is partly due to score normalization. For instance, the MAE score translates average error into a percentage against a maximum possible error (180 degrees), meaning moderate average errors still yield high percentages. The Cosine Similarity will also be high if local segments maintain some resemblance in shape and relative magnitude.

- **Limitations:** The primary limitation remains the per-joint analysis, which doesn’t capture holistic aspects like inter-joint coordination, overall posture, or movement fluidity. These unmeasured characteristics likely contribute significantly to the perceived quality differences. Another limitation is that MediaPipe relies on facial features like the eyes, nose, and mouth (which are often missing or unclear in non-human in avatars) leading to inaccurate keypoint detection. In addition, varying backgrounds and visual effects in the avatar videos can disrupt consistent pose estimation in some frames.

9 Conclusions

This project successfully developed and validated a pipeline for comparing dance movements using pose estimation and a suite of similarity metrics. Our main contribution is the effective application and comparative analysis of Windowed Cosine Similarity, MAE-based similarity, and Correlation-based similarity.

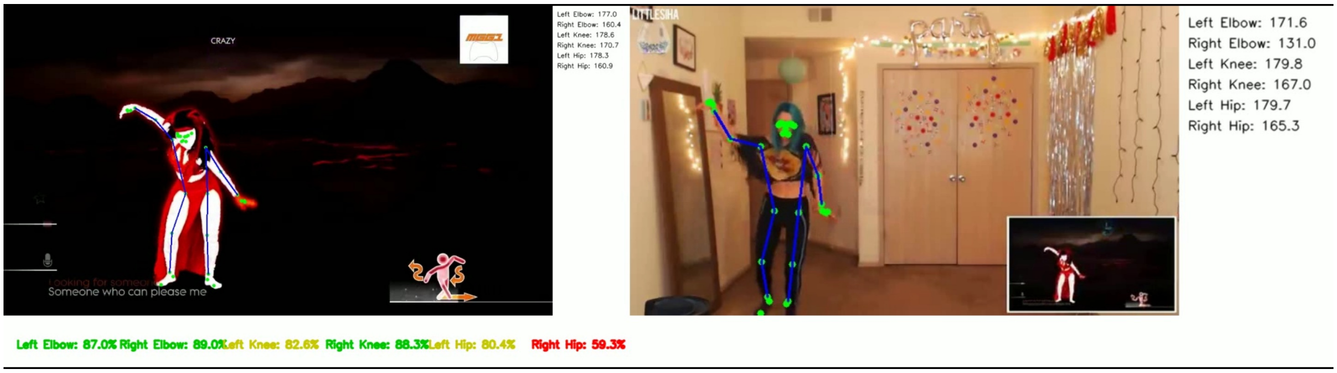
After resolving data synchronization issues, all three metrics demonstrated the ability to distinguish between ”Good” and ”Bad” human imitations when compared to an avatar reference. This study underscores the importance of selecting appropriate metrics for nuanced tasks like dance evaluation and highlights how different mathematical measures capture distinct aspects of motion similarity.

Future work could involve incorporating features that describe movement dynamics (e.g., velocity, smoothness) and inter-joint coordination. Exploring machine learning approaches trained on datasets of rated dance performances could also lead to more robust and human-aligned quality assessment.

References

[1] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Urtasun, R., Possamaí, L. A., & Grundmann, M. (2019). MediaPipe: A Framework for Building Perception Pipelines. *arXiv preprint arXiv:1906.08172*.

A Appendix: Visual Results

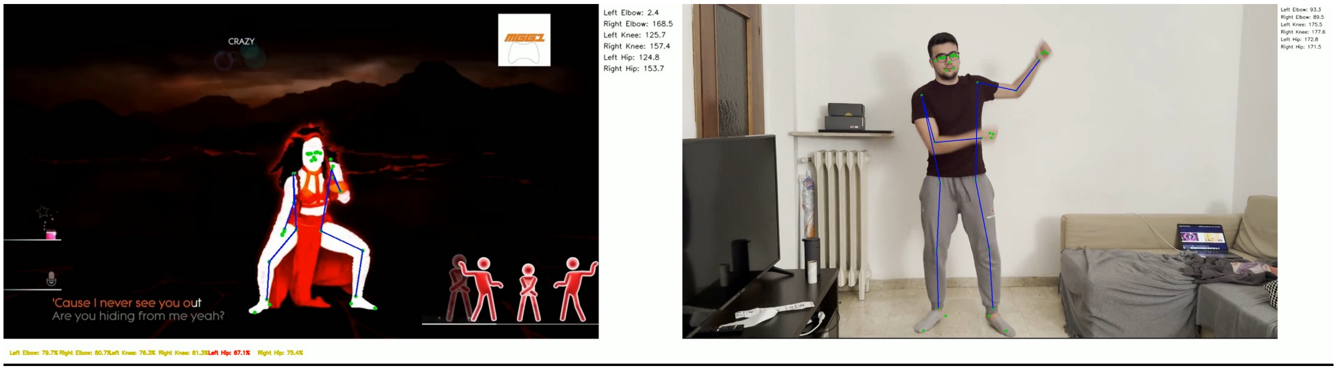


Left Elbow: 177.0
Right Elbow: 160.4
Left Knee: 178.6
Right Knee: 170.7
Left Hip: 178.3
Right Hip: 160.9

Left Elbow: 171.6
Right Elbow: 131.0
Left Knee: 179.8
Right Knee: 167.0
Left Hip: 179.7
Right Hip: 165.3

Left Elbow: 87.0% Right Elbow: 89.0%Left Knee: 82.6% Right Knee: 88.3%Left Hip: 80.4% Right Hip: 59.3%

Figure 1: Side-by-side comparison of Avatar (left) and Good Human Imitation (right) with dynamic MAE-based similarity scores.



Left Elbow: 2.4
Right Elbow: 168.5
Left Knee: 125.7
Right Knee: 157.4
Left Hip: 124.8
Right Hip: 153.7

Left Elbow: 93.3
Right Elbow: 89.5
Left Knee: 175.5
Right Knee: 177.6
Left Hip: 172.8
Right Hip: 171.5

Left Elbow: 79.7% Right Elbow: 80.7%Left Knee: 76.3% Right Knee: 81.3%Left Hip: 67.1% Right Hip: 75.4%

Figure 2: Side-by-side comparison of Avatar (left) and Bad Human Imitation (right) with dynamic MAE-based similarity scores.