Evaluating Body Pose Grouping via Clustering Algorithms

University of Hertfordshire UH

Mahendra Nath Reddy E, Priyanka Babu Rajasree, Dr. Paul Moggridge;
Corresponding author: p.moggridge@herts.ac.uk
School of Physics, Engineering and Computer Science

Code Repository

Abstract

Technologies such as MediaPipe and OpenPose have demonstrated the easy extraction of body pose data from video streams. While supervised classification models are highly effective at classifying body poses, they require massive amounts of manually labeled data, limiting scalability.

Unsupervised clustering offers a potential solution to accelerate the labeling process by grouping continuous streams of coordinates into discrete clusters. This research investigates the effectiveness of clustering algorithms (K-Means, DBSCAN, and Agglomerative clustering) for grouping raw body pose data.

We propose a preprocessing pipeline that utilizes a dual-stage heuristic filter to remove "pose hallucination" artifacts before trialing several clustering algorithms. Our results demonstrate that clustering can successfully find and extract discrete physical postures with minimal data preprocessing, offering a scalable alternative for annotation-source biomedical image and video analysis.

Introduction

Prior studies have explored the effectiveness of clustering algorithms for grouping body pose data through feature extraction and unsupervised learning techniques, often relying on latent features without directly evaluating clustering effectiveness on raw pose data. This work highlights that raw pose data contains highly discriminative spatio-temporal patterns.

Our **research question**: How effective are clustering algorithms for grouping raw body pose data? this study focuses on systematically evaluating K-Means, DBSCAN, and Agglomerative clustering on raw coordinate streams to better understand their impact on performance.

The selection of Indian Classical Dance (Bharatanatyam) as our primary data source was predicated on its high-entropy movement profile and repeatable geometric structures. Unlike casual human motion, classical dance involves discrete, predefined postures such as the Araimandi (half-squat) and Sthanaka (standing), which provide natural "ground-truth" clusters within a continuous coordinate stream.

Methodology

1. Data Preprocessing & The "Pose Hallucination" Problem

The source video (trans-coded to H.264 at 59.94 fps to ensure sub-frame alignment) was processed using the MediaPipe Pose Land marker (Model Complexity 2) to extract 33 3D landmarks (x, y, z) per frame.

While MediaPipe is computationally efficient, it exhibits a significant limitation known as "Pose Hallucination." When the camera frame transitions to a medium-shot (hip-to-head), the model continues to regress coordinates for the lower limbs based on the orientation of the torso, despite the limbs being occluded or out-of-frame.

This behavior introduces significant noise into the dataset, as the "estimated" points do not reflect the dancer's actual pose.

2. The Dual-Stage Heuristic Filter

To mitigate these hallucination errors and clean the data prior to clustering, a dual-stage heuristic filter was implemented:

- **Visibility Filtering**: We monitored the visibility probability (v) for the ankle landmarks (indices 27 and 28). Frames with ankle visibility scores below 0.65 were discarded. Qualitative review showed landmarks below this threshold were consistently associated with unstable or implausible lower-limb estimations.

- **Geometric Constraint Filtering**: We measured the normalised vertical distance between the mid-hip and the ankles. Frames where this "leg length" was less than 0.18 of the image height were identified as "collapsed skeletons" and removed.

This filtering process successfully reduced the raw dataset from 13,474 frames to a 4,789-frame high-fidelity subset, vastly improving the Signal-to-Noise Ratio (SNR) for the clustering phase.

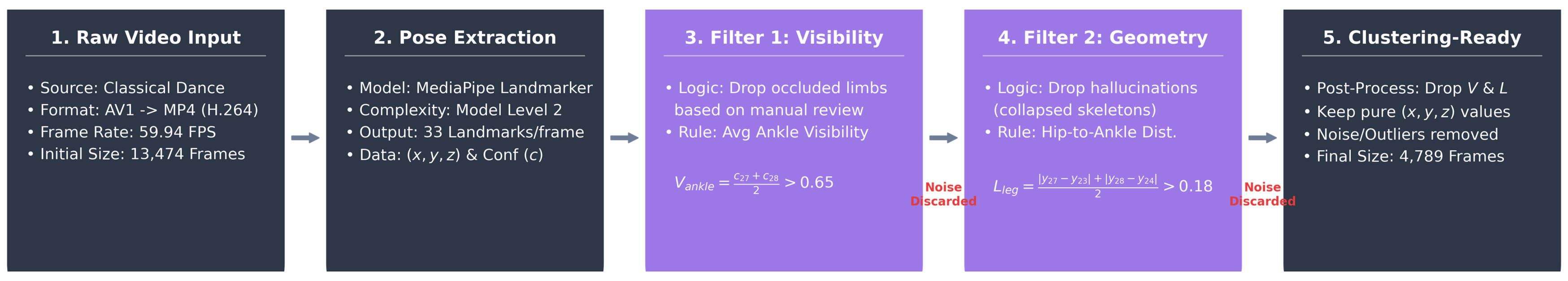

Fig: 1

3.The Clustering Algorithms

K-Means Clustering: The K-Means algorithm was used as the primary clustering method due to its scalability and performance as a baseline for unsupervised pose grouping. Before clustering, the coordinates were normalised using a StandardScaler to ensure zero mean and unit variance across the 99-feature vector (33 landmarks × 3 dimensions).

To objectively determine the number of discrete pose clusters (k). We compared the Silhouette Coefficient and the Elbow Method (using the Kneedle algorithm).

While the Silhouette score suggested (k=2) (identifying only broad movement vs. stillness), the Elbow method identified a distinct inflection point at (k=6). Given the complexity of the dance repertoire, k=6 was selected to capture the nuanced variations between specific postures.

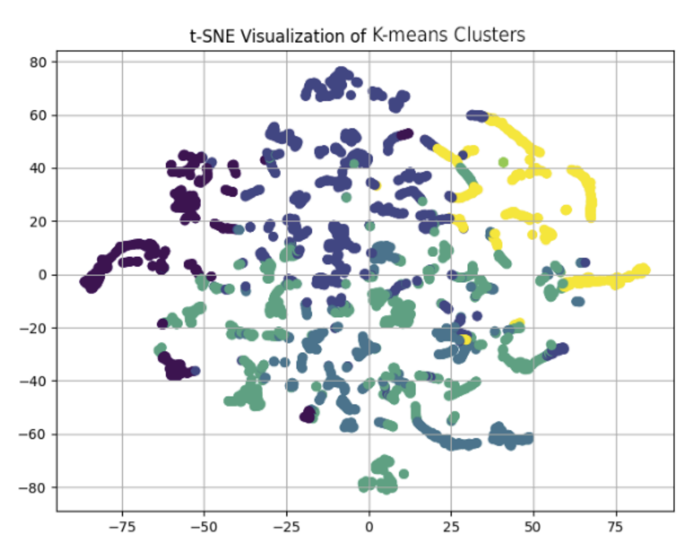

Fig: 2

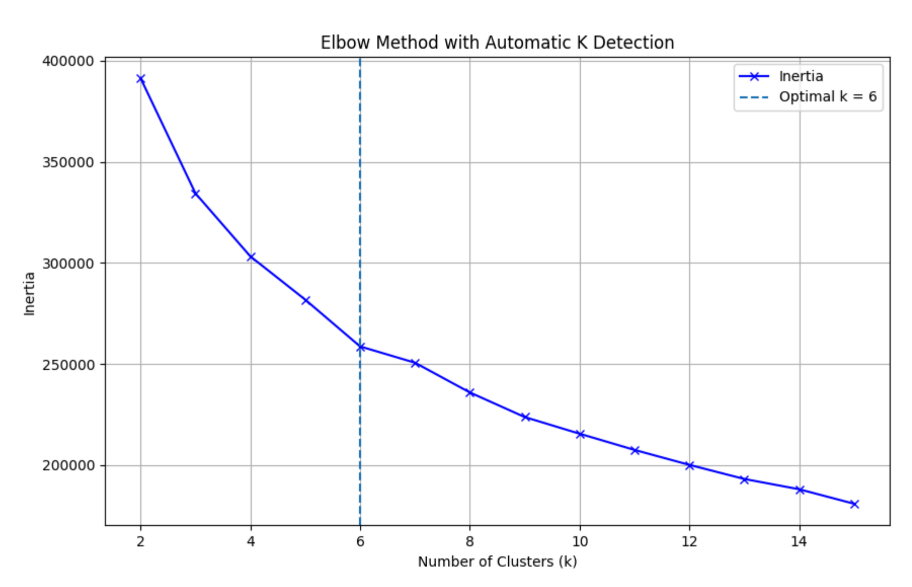

Fig: 3

DBSCAN Clustering: DBSCAN with OPTICS was applied to the StandardScaled coordinates. Figure 5 illustrates the reachability plot generated with min-samples = 50. A reachability distance threshold of 2.4 was selected to produce five distinct clusters, plus one noise cluster, based on the prominent valleys observed in the plot. This distance was specifically chosen to ensure comparability with the K-Means results.

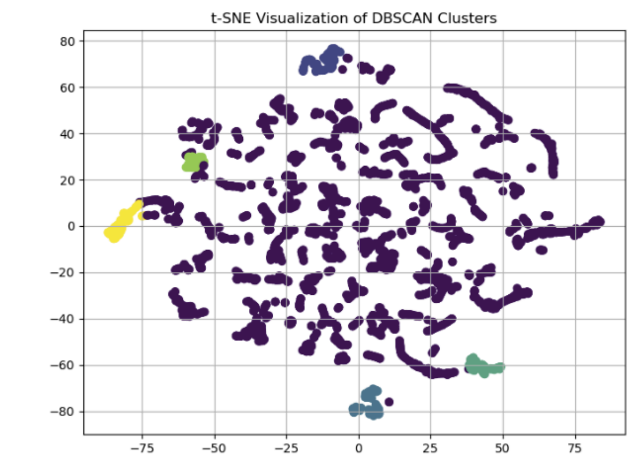

Fig: 4

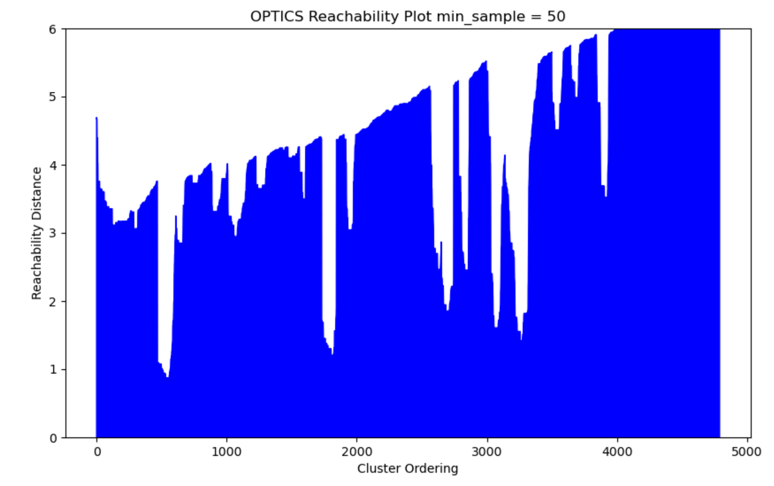

Fig: 5

A significant advantage of DBSCAN is its inherent ability to handle noise and outliers. In this context, the algorithm effectively ignores intermediary poses and transient inaccuracies in the MediaPipe coordinates. Qualitative evaluation of the 3D scatter plots in Figure 8 confirms that DBSCAN successfully identifies unique postures. Notably, clusters C1 and C4 represent similar physical poses performed at different spatial coordinates within the video frame.

Agglomerative Clustering: Agglomerative (hierarchical) clustering was performed on the scaled dataset. This algorithm trialled as it has been shown to outperform other methods [4]. To maintain consistency with the K-Means elbow point, a distance threshold of 175 was applied. The dendrogram in Figure 7 suggests stable clustering configurations at both k=5 and k=7, though k=6 remains a viable middle ground for comparison.

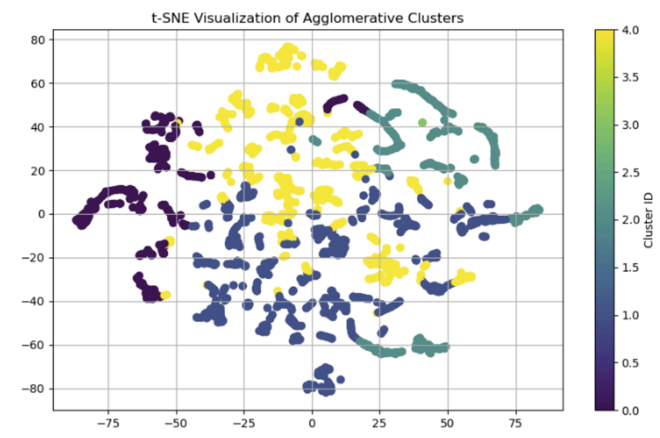

Fig: 6

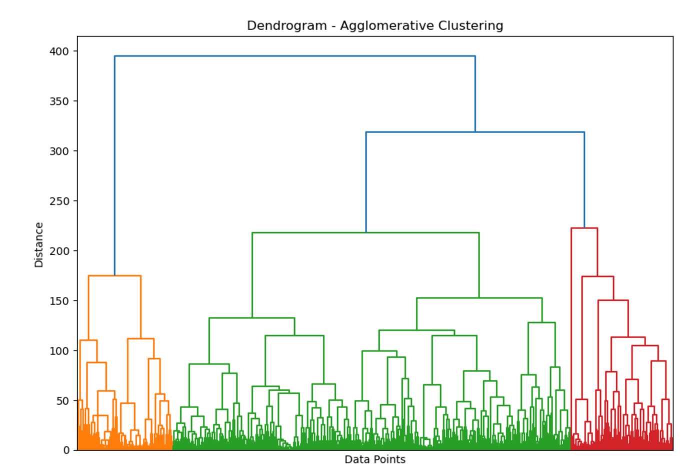

Fig: 7

Results and Discussion

Cluster Validation and Dimensionality Reduction: To validate the resulting clusters, Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) were employed for linear and non-linear dimensionality reduction. t-SNE was configured with a perplexity of 30, effectively mapping the high-dimensional 99-feature pose vectors into a 2D space. The visualisations demonstrated clear spatial separation of the 6 clusters, confirming the distinctiveness of the extracted pose classes in Fig 2, 4, and 6.

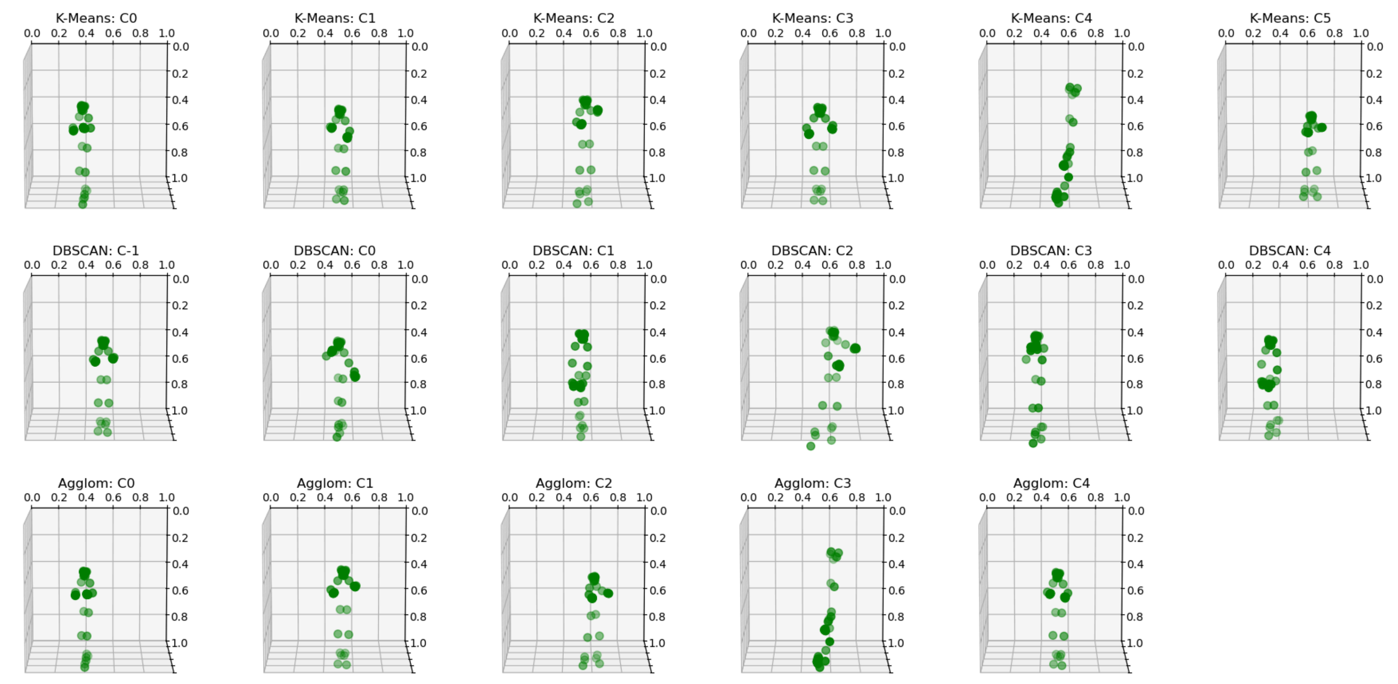

Fig: 8

3D Pose Visualisations & Observations: Qualitative evaluation of the 3D scatter plots confirms that unsupervised clustering successfully identifies unique postures. DBSCAN proved particularly advantageous; the algorithm effectively ignored intermediary poses and transient inaccuracies, isolating them strictly as noise. However, the methodology revealed a limitation in spatial normalisation: identical physical dance poses performed at different spatial coordinates within the video frame were sometimes classified as distinct clusters (e.g., Clusters C1 and C4 in the DBSCAN output) in fig 8.

Conclusion and Further Work

Conclusion:

- This investigation demonstrates that K-Means, DBSCAN, and Agglomerative clustering are highly capable of identifying and grouping discrete poses from continuous MediaPipe body landmark data.
- DBSCAN proved exceptionally effective at isolating temporary inaccuracies and transitional movements as noise, maintaining high cluster purity.
- By transforming continuous coordinate streams into discrete, manageable action groups, these unsupervised techniques can significantly accelerate the data annotation pipeline for training supervised pose classification models.

Future Work:

- Spatial Normalization: Implement spatial normalization techniques to ensure identical poses in different frame locations are clustered together.
- Advanced Features: Integrate temporal motion features (velocity, acceleration) and frequency-aware embeddings to transition from static pose grouping to full action recognition.
- Edge Deployment: Combine this clustering methodology with transfer learning to develop passive-monitoring and fall-detection systems for edge devices in elderly care.

References

[1] J. Jangade and K. S. Babulal, "Study on Deep Learning Models for Human Pose Estimation and its Real Time Application," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023, pp. 1-6, doi: 10.1109/ISCON57294.2023.10112004.

[2] S. R. Krishna, M. Pavithra, A. P. A, J. G. Thatipudi, G. D and K. Rajkumar, "MNCL: Empirical Development of Human Action Recognition Based on Postures Using Modified Neural Classification Logic," 2025 International Conference on Visual Analytics and Data Visualization (ICVADV), Tirunelveli, India, 2025, pp. 1299-1306, doi: 10.1109/ICVADV63329.2025.10961249.

[3] C. Qu, R. Chen and X. Che, "Motion Arc Analysis in Virtual Reality Environment," 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Prague, Czech Republic, 2022, pp. 3384-3389, doi: 10.1109/SMC53654.2022.9945281.

[4] W. Vexo, C. Jeenanunta, S. Chanyachatchawan, A. Tunpan and N. Sirimarnkit, "Unsupervised Human Action Recognition for Quality Control in Industrial Environments," 2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, 2024, pp. 541-546, doi: 10.1109/IICAIET62352.2024.10730750.

[5] X. Wu, P. Zhang, J. Wu, M. Wang, W. Qin and T. Liu, "Spectral Channel Attention for Skeleton Action Recognition in Lubricant Filling," 2025 6th International Conference on Control, Robotics and Intelligent System (CCRIS), Guangzhou, China, 2025, pp. 1-5, doi: 10.1109/CCRIS65764.2025.11193097.