Video Movement Detection Model

Edo Lior 203339510

edoli@post.bgu.ac.il

Software and Information Systems Engineering, Ben-Gurion University,

Beer Sheba, Israel

ABSTRACT

Dance embodies artistic expression and when technology intertwines with this form of expression, it represents a unique domain of investigation and research. Numerous studies have attempted to automate dance attributes to capture motion patterns of dancers, categorize and classify dance styles [1], as well as video dancing generations [2]. Despite recent years of concerted efforts by researchers in the dance video recognition domain, no fully capable detection model has been deployed to characterize efficiently changes in dance over time. In this paper, a video motion detector model tracks and evaluates dancing movement sequences by frames of videos by generation of multiple types of time-series datasets. The goal is to characterize and capture various dancing motion patterns and to achieve dominant representative features of dance styles during recent decades. Each dance is divided into multiple components including, unique body movements, aggregated body parts distances, recurrent movement patterns, types of movement gestures by their coordinates and more. Large datasets are generated from each dance comprised of many different types of dynamic pose changes. Analyzing and extracting relevant and dominant dancing attributes serves as a big challenge, however this model has obtained significant differences results. Ten dominant dance movement features are defined, extracted and transformed into time-series representation to the utilization of recognition in dance automation. This dance structure embedded representation is evaluated on 7 labels of decades (1950s - 2020s), each decade consisting of multiple dance videos.

1. INTRODUCTION

Dance is a widely embraced art cherished by individuals across all ages around the world. In the years between 1950 and 2020, the landscape of dance styles underwent significant evolution, reflecting shifts in culture, music, technology, and social norms. In the 1950s, dance styles such as swing, jazz and blues were popular, characterized by energetic movements and partner-based routines [3]. The 1960s brought the emergence of new styles such as the twist, influenced by rock and roll music and the development of various line dances. The 1970s witnessed the rise of disco, marked by synchronized movements and vibrant discotheque culture. In the 1980s, hip-hop and breakdance emerged as dominant dancing styles, originating from urban street culture with its unique style and dance moves. Techno genre originated in the 1980s from the United States, as a response to the city's industrial decline. Techno is characterized by its synthetic sounds, repetitive beats, and futuristic themes, often created using synthesizers, bass and drums. The 1990s saw the globalization of dance styles, with the proliferation of music videos and the spread of Latin dances such as salsa and reggaeton. Additionally, technological advancements enabled the creation of new dance forms, such as dubstep and EDM inspired choreography. The genre of trance emerged in the early 1990s, primarily in Europe, with its roots traced back to Germany. It gained popularity in clubs and raves, characterized by its repetitive melodic phrases and energetic rhythms. Trance music was heavily influenced by other electronic music genres such as techno, house, and ambient music. In the 2010s, dance styles also became influenced by social media platforms with choreographers blending elements from various genres to create new and innovative movements [4]. This fusion was particularly evident in contemporary dance, where choreographers drew inspiration from diverse cultural traditions, music genres, and artistic disciplines to create new dance performances.

Overall, the time period between 1950 to 2020 witnessed a dynamic evolution of dance styles, reflecting changes in society, music, technology, and artistic expression. From the energetic partner dances of the mid-century to the global fusion and digital dissemination of contemporary trends, dance remains a vibrant and evolving art form, reflecting the choreography and spirit of its time.

Each dance is consisted of various data types from poses, action recognitions, body movements, to audio beats and instruments that create an entire range of diverse dance styles. Time-series temporal based methodologies and Deep learning architectures have been used for their efficacy in addressing diverse challenges in fields ranging from frame processing and movement recognition to video style classifications [5]. Machine learning predicting models that are temporal based have also been proposed. These models enable us to figure out what movement patterns are derived from a single dance or a whole decade of dances. The goal is to extract the most frequent and relevant relations of patterns that can represent a detailed embedded structure of a dance. The data collected are based on intervals of timestamps from dancer's measurements of poses in each frame of the video. These patterns are used as features that represent unique or recurrent dance movements. Algorithms based on higher resolution of the time variable assist in defining recognition attributes in a higher detail of information. In this study, a video detector model is implemented to detect multivariate data from each dance video. In addition, the model generates and evaluates time-series datasets based on movement recognition automation.

2. RELATED WORK

2.1 Deep Learning Movement Recognition

Usage of deep learning approaches have solved complex problems in video classification tasks. In the pursuit of advancing traditional methods for video content identification, novel techniques have emerged. Among which is the framework proposed by Bakalos, et al. [6] for recognizing and structuring dance poses within videos. Utilizing Kinect sensors for RGB capturing and real-time depth sensing, this method captures authentic postures, drawing a sequence of poses detected by the Kinect sensor and assigning appropriate labels to each frame. Similarly, Shuhei Tsuchida conducted a study using the same database, presenting four baseline methods for classifying dance genres employing LSTM and SVM classifiers [7]. These methods include the adaptive method, where beat positions are treated as single units corresponding to video frames, and the L-fixed method, wherein vectors accumulate within fixed-length units. Results showed that the L-fixed method with LSTM model achieved the highest accuracy of 91.4%. Furthermore, Maale and Pushpanjali [8] demonstrated the efficacy of Support Vector Machine (SVM) classification when combined with the Shannon Entropy Algorithm, achieving significant results. Specifically, their proposed algorithm, when integrated with an SVM classifier, achieved 94.4% accuracy compared to the existing neural network model, which reached 74.8% accuracy. The potential applicability of this algorithm extends to real-time dance video classification and can be further adapted for classifying unlabeled dance styles.

2.2 Time Series Movement Recognition

The data being used consists of multivariate temporal data each based on different time lengths of samples and different types of movement actions. Data mining analyzes large volumes of unstructured data to automatically discover interesting regularities to a better understanding of the collected data. The goal is to discover recurrent patterns that are most relevant for determining dominant movements. It can be used in various fields such as: financial predicting systems in businesses, stock markets, bank loans, and insurances. More fields use it for scientific experiments, warning systems for ecological disasters like earthquakes, dangerous weather and oil leaks. Each record of data is referred to as a timestamp. A time series is a collection of timestamps $T = \{t_1, t_2, t_3, \ldots, t_n\}$ which is an ordered set of n measurements taken over a period of time and can be represented by different scales, from nanoseconds to years. A study introduced a real-time system for classifying gestures depicted in skeletal wireframe motion [9]. Its core elements consist of an angular skeleton representation, tailored to ensure robust recognition amidst noisy input and a cascaded correlation-based classifier for handling multivariate time-series data. They evaluated 28 gesture classes, with hundreds of instances recorded using the XBOX Kinect platform and performed by multiple subjects for each gesture class. Their results show classification performance of 96.9% accuracy for approximately 4-second skeletal motion recordings. This remarkable accuracy is noteworthy considering the presence of input noise arriving from the real-time depth sensor.

Another study introduced the task of dance beat tracking which is a fundamental area within Dance Information Retrieval (DIR) research [10]. It involves identifying musical beats within a dance video solely through visual cues, excluding any reliance on music audio information. This visual analysis of dances is crucial for achieving a comprehensive machine understanding of dances, extending beyond those accompanied by music. As a subset of Music Information Retrieval (MIR) research, DIR shares similar objectives with MIR and seeks to extract various high-level semantics from dance videos. Despite extensive research on audio-based beat tracking within MIR, there has been limited exploration of visual-based beat tracking for dance videos. The data repository used was the AIST Dance Video Database [11]. This database includes 13,888 music videos in MP4 and WAV format of 10 dance styles. Time series classification performances demonstrated that Temporal Convolutional Neural Networks (TCNs) outperform significantly bidirectional LSTMs in accuracy.

A recent paper introduced a Body Sensor Network Model based on a Deep CNN architecture that detects dance activity recognitions of autonomously learning distinct morphological features from sensor data and utilizing them to accurately identify dance steps [12]. Their model achieved 93.53% accuracy while extracting 46 features from time windows and comparing with the following models, Naive Bayes, KNN (k=5), Linear SVM, Multilayer Perceptron and Random Forest. The data collected was from dance routines from Indian classicals performed by a professional dancer. They encountered challenges posed by heterogeneity and sampling rate instability across different sensor modalities in body sensor networks. They presented a solution involving a video recording Based ground truth data annotation synchronizer, enabling precise labeling of extensive dance activity datasets. In this study, time series based on video frames are analyzed and the values that are measured are by landmark poses of the body.

3. METHODS

3.1 Preprocessing

Dancing video files were downloaded using package pytube [13]. The videos contain multiple dancing styles defining a decade. Each video file was divided into 4 experimental window sizes of: {1, 3, 5, 10}.

3.2 Data Imbalance

Since each label of a decade contains a different amount of video observations, each movement feature is normalized and averaged according to the number of videos in a certain decade.

3.3 Movement Recognition

To detect a list of all important joints that contribute to different kinds of movements, this study uses Google’s 3D human body recognition framework MediaPipe [14, 15, 16, 17, 18]. This package consists of built in sub models that show the relationship between all the landmark components to classify the connection between different body parts to predict human body pose and emotions. This tool detects 32 key points of parts of the body. For higher recognition efficiency, it calculates normalized landmark coordinates that represent a point in 3D space with x, y, z coordinates. x and y are normalized to [0.0, 1.0] by the image width and height. To detect the location of the recognitions in the frame, the normalized coordinates are multiplied by the frame’s width and height accordingly.

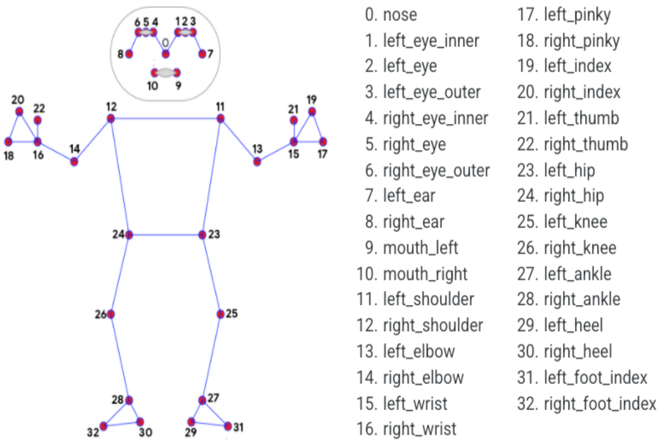


Fig 1. MediaPipe 32 Determining Joints

3.3.1 Defining Dance Tracking Locations

Dance features can represent a single dance, multiple dance styles or a dance style defining a decade. It is a complex structure to represent, thus dance movements by all hierarchal levels of the human body are captured to find the most essential dominant dancing patterns. Definition of tracked recognitions:

Level 1:
Atomic Key Points = *{nose, shoulder, elbow, wrist, hip, knee, ankle}*

Level 2:
Arm = *{shoulder, elbow, wrist}*
Leg = *{hip, knee, ankle}*

Level 3:
Head = *{nose}*
Upper Body = *{left arm, right arm}*
Lower Body = *{left leg, right leg}*

Level 4:
All Body = *{head, upper body, lower body}*

3.3.2 Defining Dance Unique Moves

Movements are defined by calculating angle changes between frames and tracking its state and coordinates. Definition of unique moves:

Direction Definitions

Left = left hand of the detected object. From observer point of view, it is located in the right side of the image.

Right = right hand of the detected object. From observer point of view, it is located in the left side of the image.

Left Facing Left = left hand of the detected object pointing to the left side.

Left Facing Right = left hand of the detected object pointing to the right side.

Angle Definitions

1. Left/right arm up:

a) Left arm facing left: above 30 degrees.
b) Left arm facing right: under 180 degrees.
Right arm up degree was set according to the left arm threshold degree.

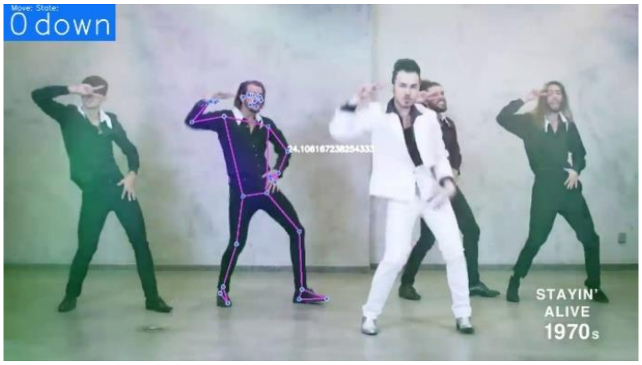


Fig 2.1.1 Left arm (facing right) state down (<30 degrees)

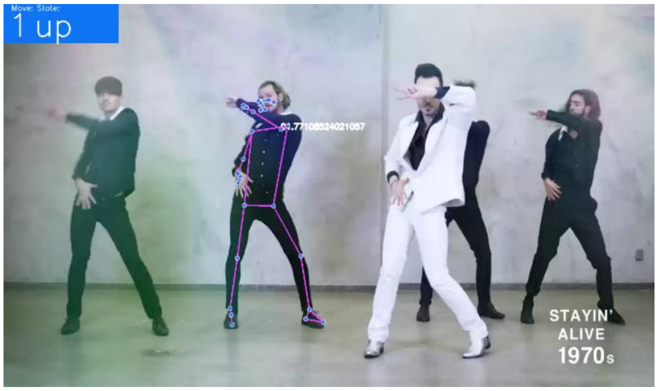


Fig 2.1.2 Left arm (facing right) state up (<180 degrees)

2. Left/right leg up:

a) Left leg facing left: under 160 degrees.
b) Left leg facing right: under 165 degrees.
Right leg up degree was set according to the left leg threshold degree.

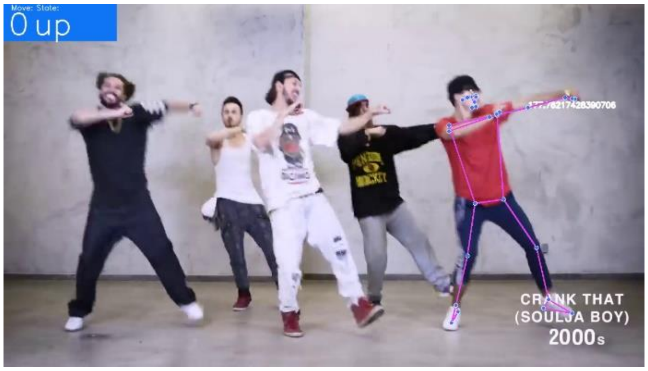


Fig 2.2.1 Left leg (facing left) state down (<160 degrees)

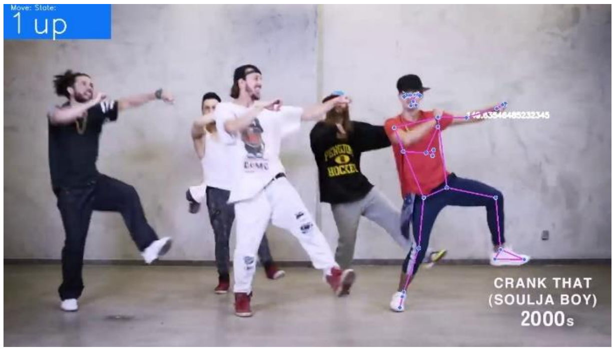


Fig 2.2.2 Left leg (facing left) state up (<160 degrees)

3. Jump:

Nose height < 160 pixels of frame height (Y axis of image frame is ascending and not descending).

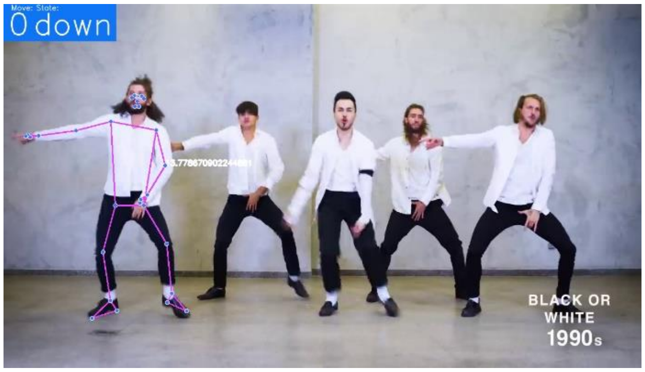


Fig 2.3.1 Jump state down (>160 pixels)

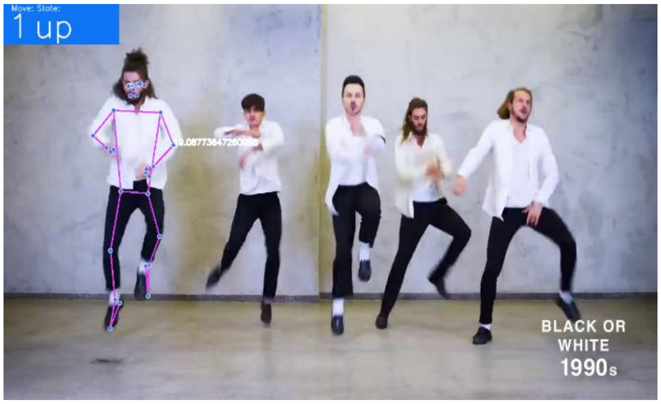


Fig 2.3.2 Jump state up (<160 pixels)

1. **4. Duck:**

Nose height > 330 pixels of frame height (Y axis of image frame is ascending and not descending).

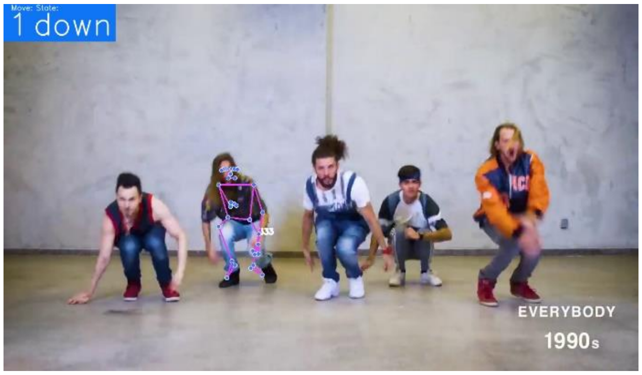


Fig 2.4.1 Duck state down (>330 pixels)

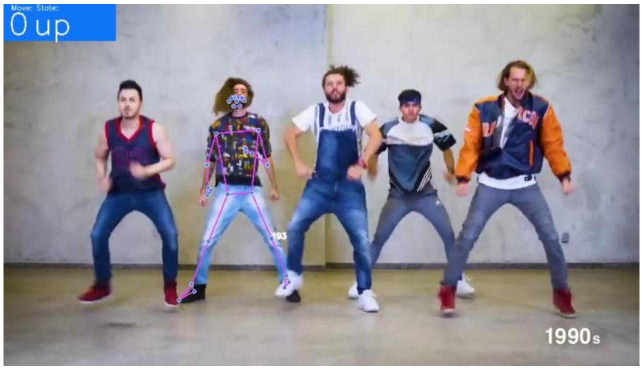


Fig 2.4.2 Duck state up (<330 pixels)

2. **5. Left/right arm down:**
Calculated according to the up arm movement degree threshold.

3. **6. Left/right leg down:**
Calculated according to the up leg movement degree threshold.

3.4 Time Series Analysis

After extracting the data from all video frames, we generate the following time series datasets for further evaluation:

3.4.1 Aggregated Distances

3.4.2 Delta Distances

3.4.3 Unique and Frequent Moves

3.4.4 Movement States

3.4.5 Pose Landmark Coordinates

Each embedded movement recognition time-series dataset is divided to 3 hierarchical timepoints by:

1. Frames
2. Videos
3. Decades

These time series datasets first apply temporal abstraction. First stage of this process is discretizing the quantitative data and representing the transformation of ambiguous time series observations into explanatory intervals of symbols. It does so by generating cutoffs that represents bins or states. These states allow us to abstract enormous amounts of observations to meaningful diversions while lowering the volume of the data. The goal is to abstract the data into higher-level concepts useful for properly recognizing all aspect details of the dance. It also helps understand how changes in movement counts and patterns evolve so it is possible to observe its unique characteristics. In this study, the following unsupervised discretization methods were experimented, Equal Frequency Binning (EFB) and Symbolic Aggregate Approximation (SAX). EFB cutoffs are found by dividing all the values into equally sized bins, each containing the same number of samples. In the SAX method, cutoffs are found by creating a gaussian distribution from the input values. This method uses the lower bounding property to enable dimensionality reduction. Once symbolic time intervals are obtained, we next detect frequent patterns to enrich the data in a deeper resolution, time dependent. Movement states are multivariate and are categorized in to 2 bins representing both up and down movement values. The data gathered often consists of enormous amounts of timestamps which have very similar values that are unnecessary and don't promote any further essential information. Redundant data will lead to extra run time and complexity. The order of events and movements are based on relations between each pair of symbols that we've extracted from the temporal abstraction stage. These relations define and add important information for the dance embedded structure. In this study, TSFresh [19] package is applied. It provides systematic time-series feature extraction by combining established algorithms from statistics, time-series analysis, signal processing, and nonlinear dynamics with a robust feature selection algorithm. Time-series data points are interpreted in the broadest possible sense, such that any types of sampled data or even event sequences can be characterized.

4 Experiments

4.1 Data

Files: 69 dance videos files.
Labels: 7 labels of decades:
1950-1960: 4,
1960-1970: 4,
1970-1980: 9,
1980-1990: 10,
1990-2000: 10,
2000-2010: 15,
2010-2020: 17.

4.2 Research Questions

Which set of embedded movement features data will best define and capture the most detailed information of the input dance structure and to represent and associate it to its period of time?

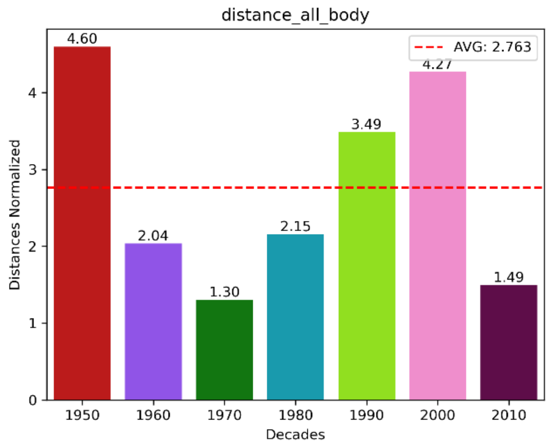
a) Aggregated distances
b) Delta distances
c) Unique moves
d) Recurrent moves
e) States of poses
f) Location coordinates of gesture recognitions

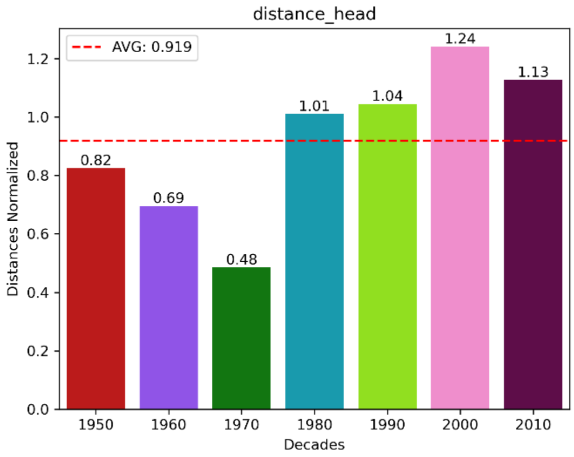
5. RESULTS

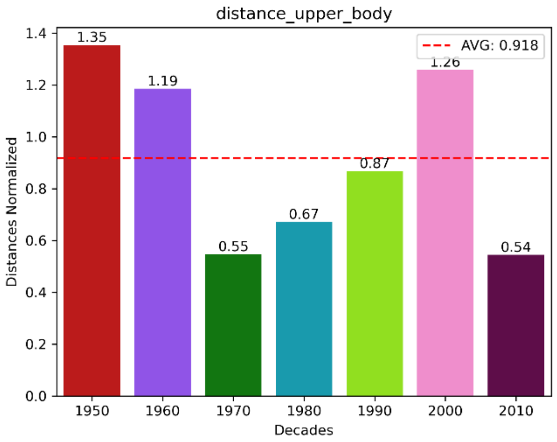
5.1 Sliding windows: movement recognitions were best represented by moving time windows of 5 frames.

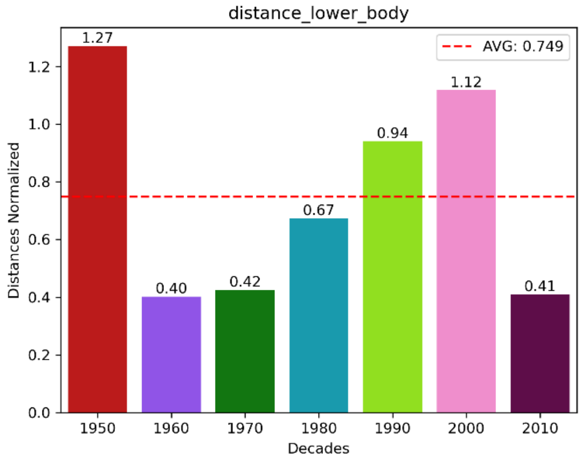
5.2 Binning: Equal frequency binning resulted in the most efficient division of time-series observations.

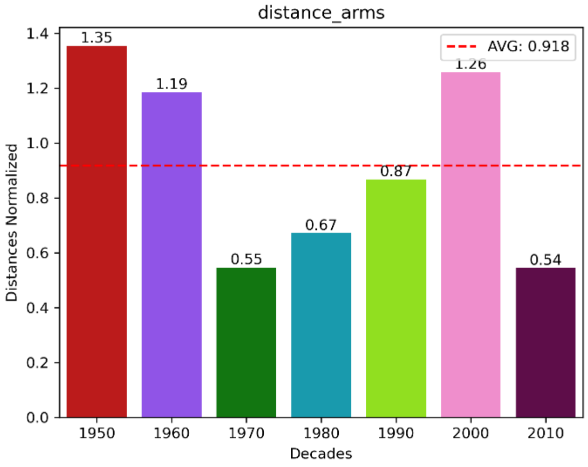
5.3 Distances:

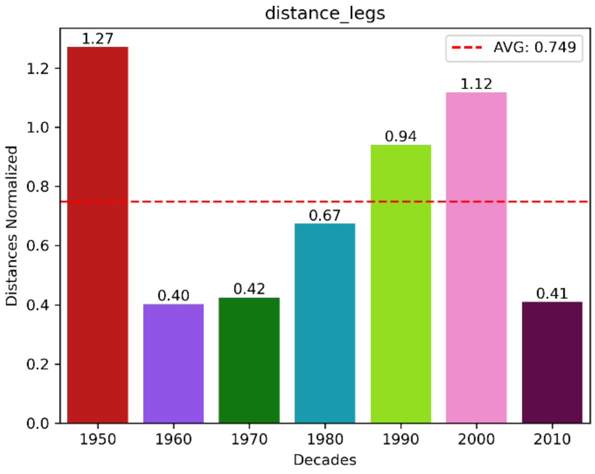


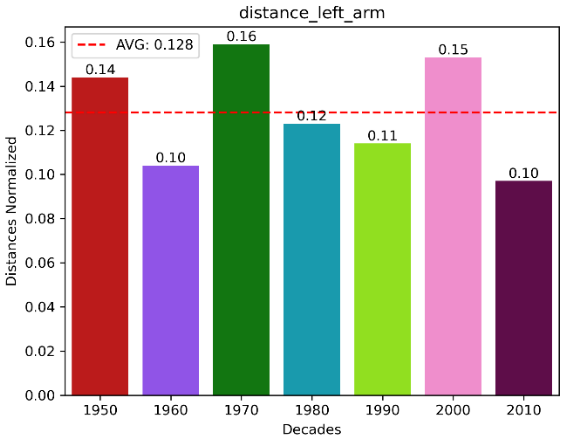


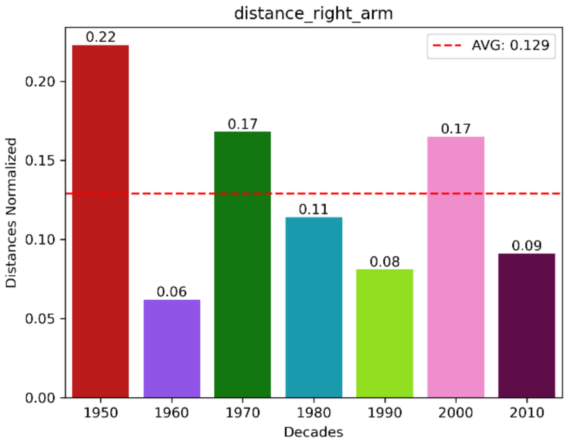


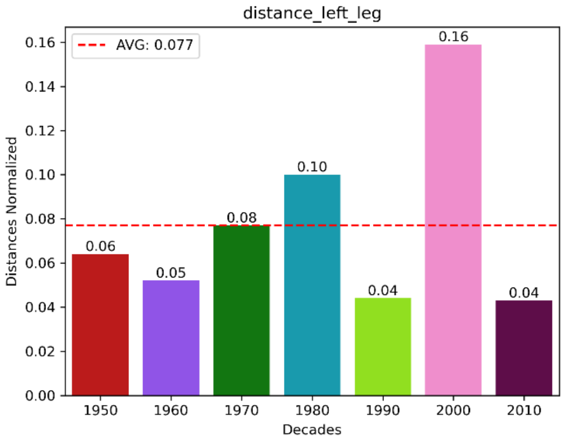


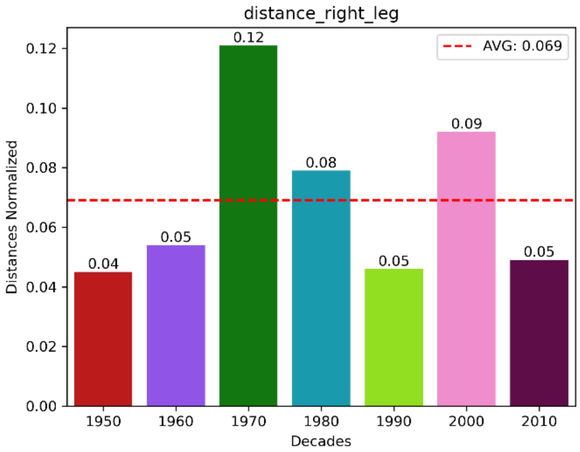








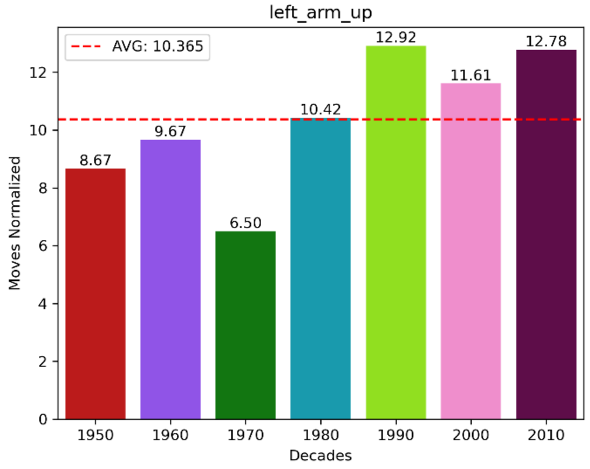


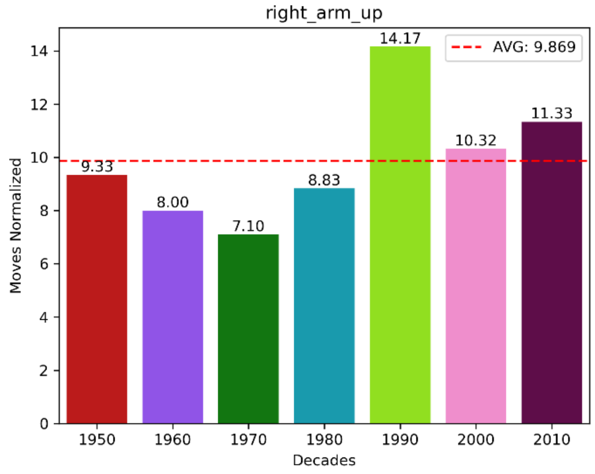


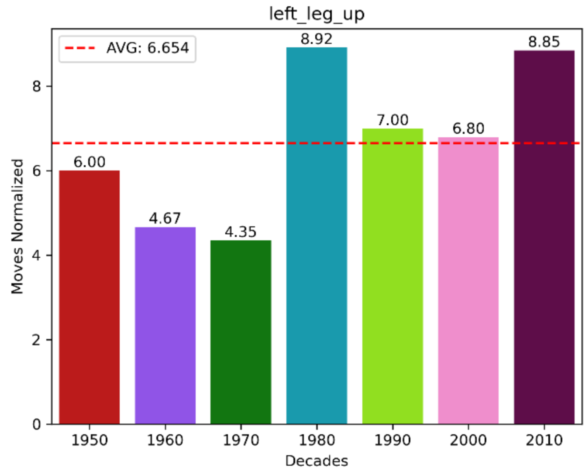
Aggregated distances of hierarchical level 4, representing all data coordinates of the entire body, in decades 1950s, 1990s and 2000s had the top distances above the decade's average. In the 1950s, the summary distance of a dancer is double the distance of dancers in the 1970s, 1980s and 2010s. The 1970s and 2010s had the lowest averaged aggregated distance measure with the whole body. Hierarchical level 3 representing the head, upper and lower body movements, indicate that the upper body and lower body aggregated distances were the highest in decades 1950s and 2000s and double the distance length than decades 1970s, 1980s and 2010s. Whereas the least distance measures were in decades 1970s, 1980s and 2010s. On the other hand, head movements were highest in decades 1980s – 2010s. Least average head movements took place in the 1970s. Hierarchical level 3 consisting of both arms and legs are highest in decades 1950s and 2000s.

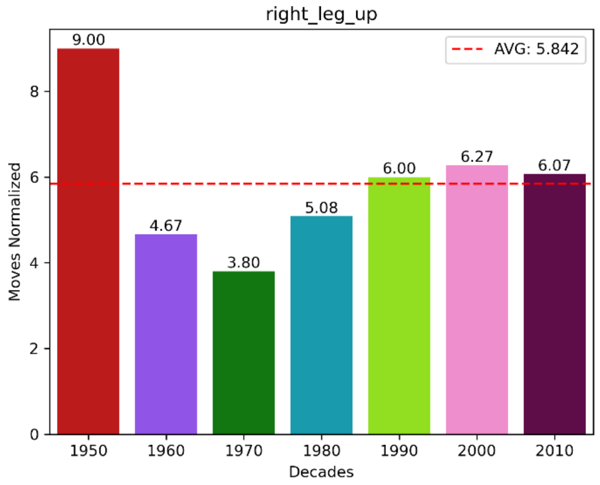
The decade of 1960s had 3 times the distance of arms than legs. In measured distances of hierarchical level 2, separating between left and right arms, all decades were persistent except for decade 1950s. In this decade the right arm travelled a significant distance longer than the left arm. However, in the 1960s the left arm travelled double the distance than the right arm. As for separating between left and right legs, all decades were persistent except for decade 1970s which had significantly longer distances with the right leg rather than the left leg. Overall, the decade of 1950 – 1960 resulted in the longest aggregative distance measured and the decade 2010 – 2020 had the shortest aggregated distance.

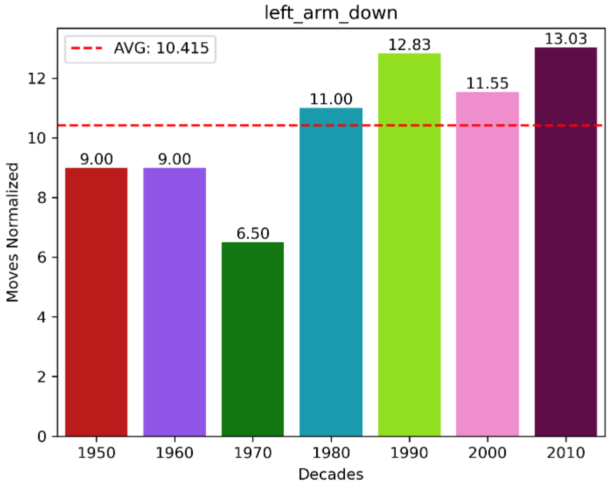
5.4 Unique Moves:

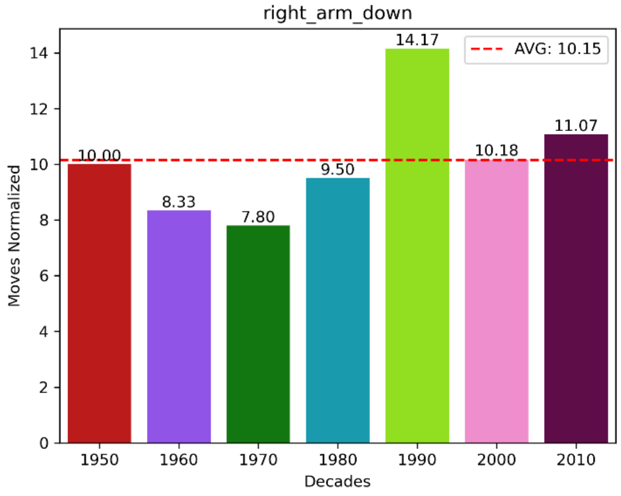


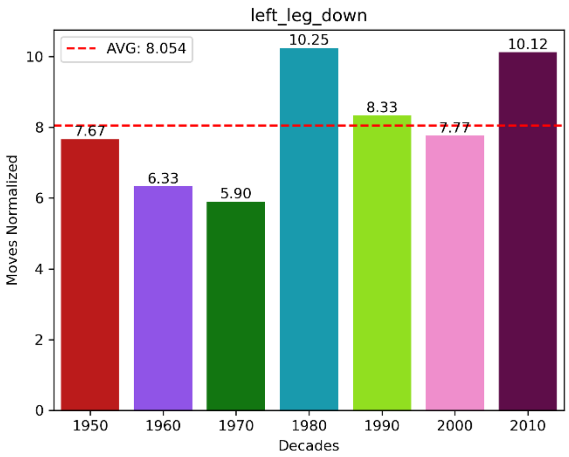


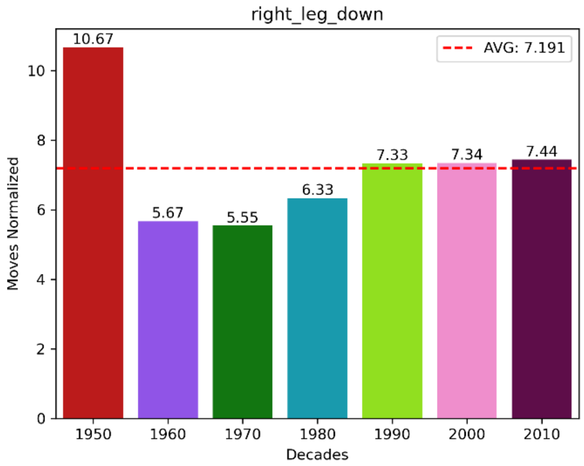


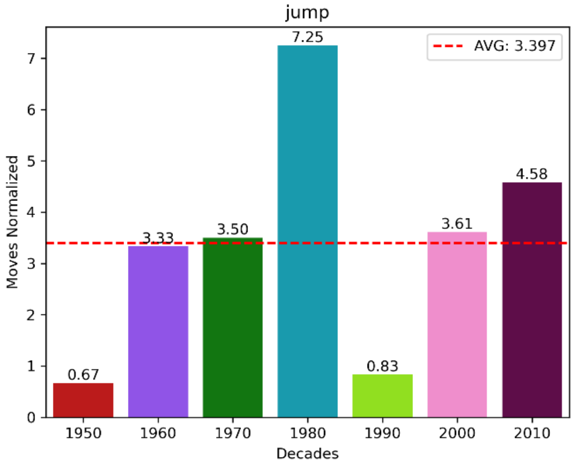


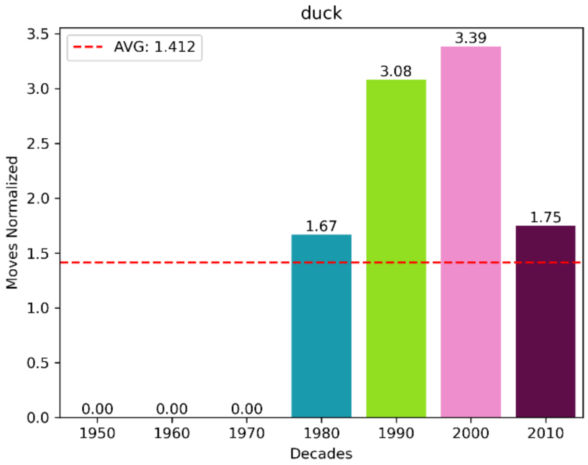












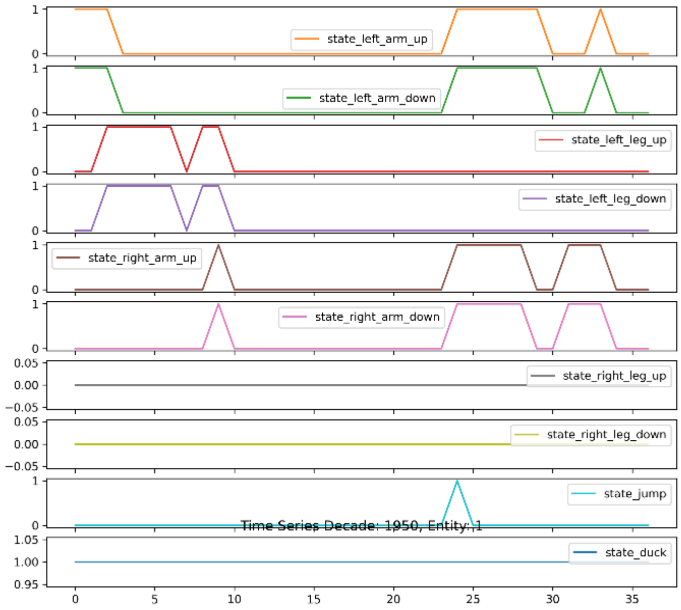
The decades unique moves representation of both left and right arms and legs rising up, was most frequent from the 1980s until the 2010s. The 1970s had minimum frequency of arms raising up. There are no significant differences between left arms and right arms rising up. In decades 1950s an 1980s there are more left leg movements rising up than the right leg rising up. However, rest of the decades have a similar ratio.

The unique action move of jumping does not exist in decades 1950s and 1990s, however the decade 1980s resulted with the most frequent count. Rest of the decades tend to have an average amount of jumping. Ducking during a dance was not featured until the 1980s – 2010s.

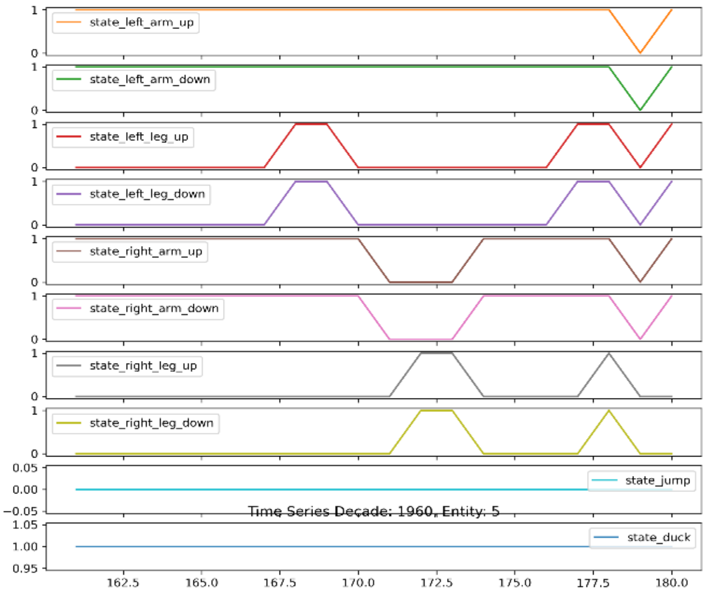
4.2 Time Series Analysis

Decade Representatives:

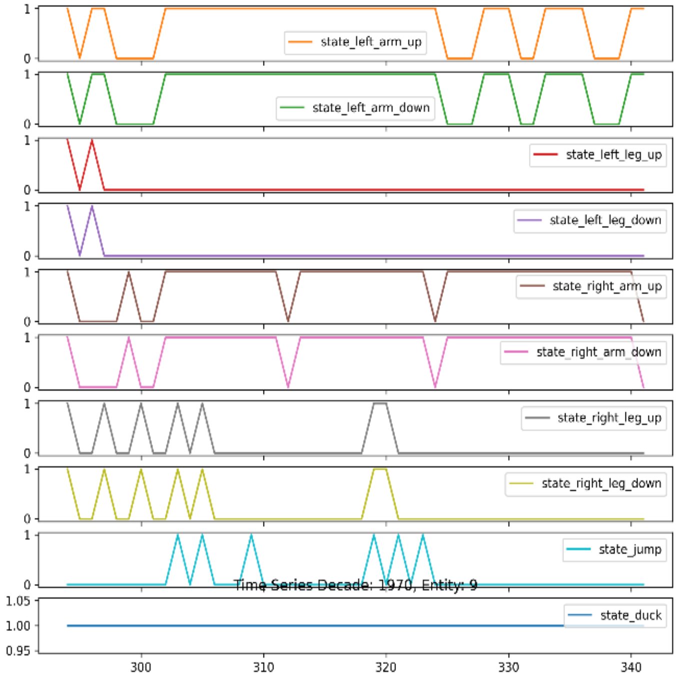
Decade: 1950, Entity 1:



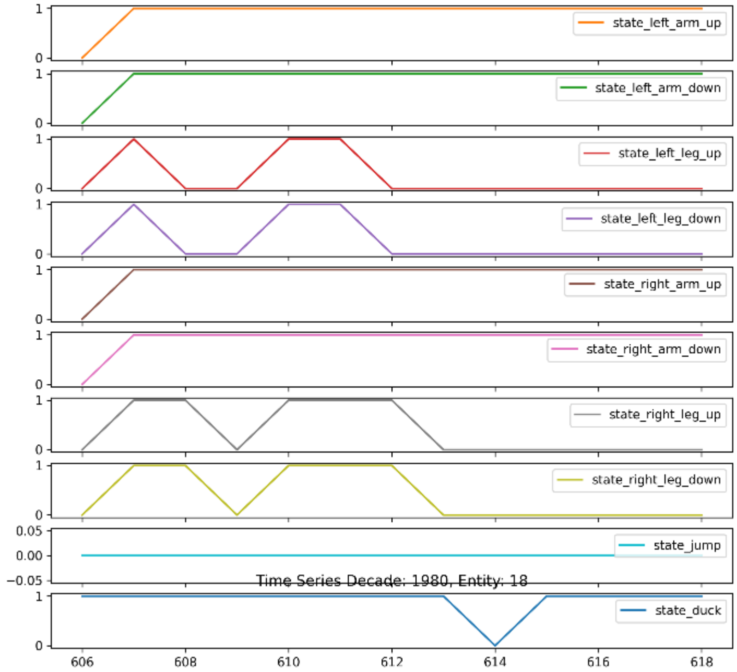
Decade: 1960, Entity 5:



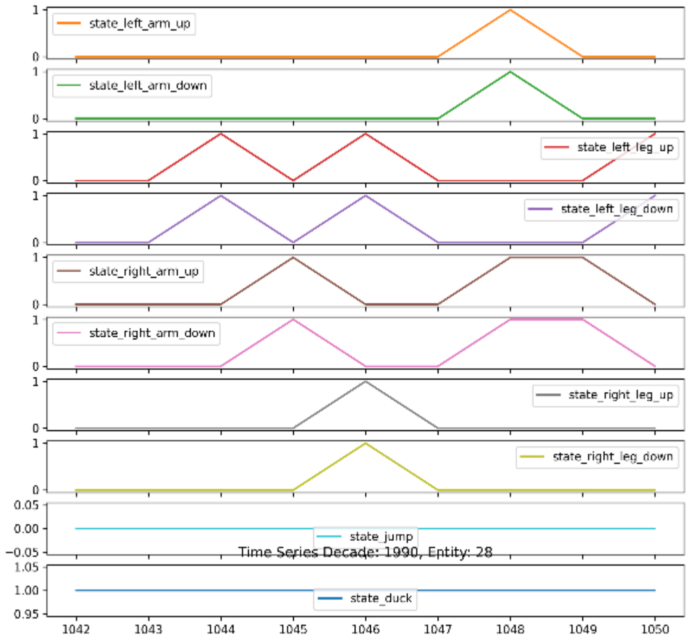
Decade: 1970, Entity 9:



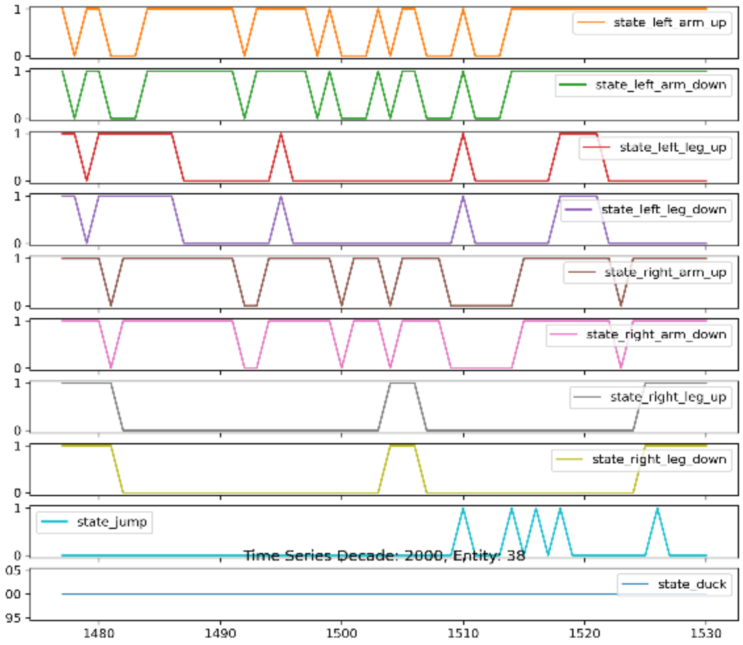
Decade: 1980, Entity 18:



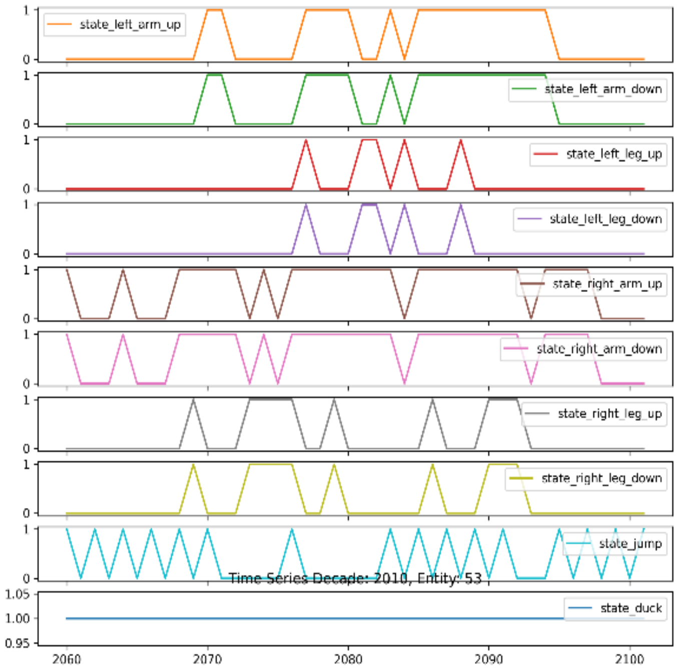
Decade: 1990, Entity 28:



Decade: 2000, Entity 38:



Decade: 2010, Entity 53:



Each figure plots a time-series structure of movement state changes. Each plot is a representative dance entity from a certain decade. These structures represent movements in time to extract dominant patterns that can represent the dance's style embedded structure.

The relations between unique movements defines time interval related patterns that represent unique and recurrent moves. This expands the data retrieved for detecting and defining dance patterns.

5.1 DISCUSSIONS AND CONCLUSIONS

Categorizing a decade of dances by the aggregated distance measure did not have significant differences in between consecutive decades, but there is a significant difference between the first decade of the 1950s and the last decade of 2010s. These results may imply that in long term of decades, body movements reduce distances and can be a cause of a new technology era including social media style dances. Categorizing dances by unique and frequent moves had a significant impact from the 1980s until the 2010s. These decades had more unique and frequent moves as opposed to decades before the 1980s. This can explain the unique transfer of recurrent movements caused by a universal shift of choreography. New dance styles with higher bpm evolved during the decades leading to higher count of moves per dance in the data. Results further demonstrate the action of ducking in dances which started to take place in the 1980s. This reason can be originated from the fact that new dance styles such as hip-hip and breakdance were introduced in those periods of time. Jumping moves were becoming common from the 1980s and has gradually grown since that time.

When applying a deep learning model for automatic feature recognition it has the ability to automatically learn relevant features from raw data, eliminating the need for pre-defined feature extractions. This is particularly beneficial for complex tasks like body movement recognition, where manually designing features can be challenging. CNNs and RNNs capture complex patterns and relationships within the data, leading to higher accuracy in classification tasks. These models can handle diverse types of body movements and variations in dancing styles. However, applying time-series based models for pattern movement detections require less data, less computational resources and most importantly are very highly interpretable for the detection of dancing attributes to help associate dominant representations of dances.

For future work, expansion of the video dataset will lead to better representation and characterization of dance styles through time. Expanding the data repository with additional decade labels will add information and will result in a higher statistical confidence for differentiating between dances. Enriching each label with various dance styles will help capture a larger amount of dominant dance motion patterns.

Examining audio recognition of the videos and extracting beats and tempo data will help determine the dance music to its decade. It is interesting to explore what detections of instruments during a dance video imply on the decade being played. An additional dataset of WAV files of music by each decade can facilitate in determining the characteristics of a video to its target decade. Experimenting dynamic window sizes dependent on tempo and beats of the input melody of the music video, will enhance and improve detection of dancers moving patterns since dance movements are built upon the rhythm. Expanding the dataset with text data of lyrics from a large repository of songs, will enable higher robustness of the model to predict the decade associated to the input dance. It is also interesting to find the most representative and popular terms of all lyrics for each decade while analyzing and investigating how terms in music changed over time. Enriching the training of the dance video detector model can also be done by image processing of the dancers’ clothes. There are numerous unique outfits for dances and can be associated to a certain time throughout the decades. Identifying dancers’ gender can also help capture dance characteristic associated to a certain decade of dance styles. Applying a multimodal video model will significantly enhance the detail of the dance embedded representation and accuracy for associating a dance representation to its decade.

5 REFERENCES

[1] Manish Joshi and Sangeeta Chakrabarty, An Extensive Review of Computational Dance Automation Techniques and Applications. Proceedings A, 2021.

[2] Mengyang Feng, Jinlin Liu, Kai Yu, et al. DreaMoving: A Human Video Generation Framework based on Diffusion Models. arXiv, 2023.

[3] Most Popular Dances History. https://www.clistudios.com/dance-blog/dance-basics/most-popular-dances-of-each-decade/

[4] Dancing Through the Decades Evolution of Trendy Dance Moves. https://dancevance.co.uk/blog-detail/dancing-through-the-decades-the-evolution-of-trendy-dance-moves

[5] Soumitra Samanta, Pulak Purkait and Bhabatosh Chanda. Indian Classical Dance Classification by Learning Dance Pose Bases. IEEE, 2012.

[6] Nikolaos Bakalos, Eftychios Protopapadakis, Anastasios Doulamis and Nikolaos Doulamis. Dance Posture Steps Classification Using 3D Joints from the Kinect Sensors. IEEE, 2018.

[7] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki and Masataka Goto. AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing. Society for Music Information Retrieval, 2019.

[8] Maale, B. R. and Pushpanjali. Recognition and Classification of Various Dance Forms Using SVM. International Journal of Research in Advent Technology, 2019.

[9] Michalis Raptis, Darko Kirovski and Hugues Hoppe. Real-Time Classification of Dance Gestures from Skeleton Animation. ACM, 2011.

[10] Fabrizio Pedersoli,and Masataka Goto. Dance Beat Tracking from Visual Information Alone. ISMIR, 2020.

[11] AIST Dance Video Database. https://aistdancedb.ongaaccel.jp/

[12] Abu Zaher Faridee, Sreenivasan Ramamurthy, et al. HappyFeet: Recognizing and Assessing Dance on the Floor. arXiv, 2018.

[13] Pytube. https://github.com/pytube/pytube.

[14] Mediapipe. https://github.com/google-ai-edge/mediapipe

[15] Amritanshu Kumar Singh, Vedant Arvind Kumbhare and K. Arthi. Real-Time Human Pose Detection and Recognition Using MediaPipe. Springer Nature Singapore, 2022.

[16] Haodong Chen and Ming C. Leu, Advancements in Repetitive Action Counting: Joint-Based PoseRAC Model with Improved Performance. arXiv, 2023.

[17] Yiqiao Lin, Xueyan Jiao and Lei Zhao. Detection of 3D Human Posture Based on Improved Mediapipe. Journal of Computer and Communications, 2023.

[18] Tianyu Wu, Shizhu He, Jingping Liu, et al. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. IEEE, 2023.

[19] Maximilian Christ, Nils Braunb, Julius Neuffer and Andreas Kempa-Liehr. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests Nerucomputing, 2018.