Video Segmentation Pipeline For Co-Creative AI Dance Application

John Gunerli
Georgia Institute of Technology
Atlanta, USA

Manoj Deshpande
Georgia Institute of Technology
Atlanta, USA

Brian Magerko
Georgia Institute of Technology
Atlanta, USA

ABSTRACT

This article introduces a method that combines human input and computation for analyzing human motion from video recordings, specifically for capturing dance movements. The central aim is to develop an innovative system for processing and analyzing videos. This system consists of four key stages: using pre-trained MediaPipe models for interactive image segmentation, organizing videos efficiently through batching, identifying and extracting keyframes, and pinpointing accurate timestamps of keyframes. This pipeline is a part of LuminAI, an interactive installation that features a virtual AI agent capable of improvising movements in collaboration with human participants. In particular, the proposed pipeline fits into the first software module of LuminAI, responsible for recognizing and segmenting continuous motion capture data into separate body actions. The proposed video segmentation pipeline is designed to fulfill the requirements for both qualitative and quantitative analyses in designing systems that classify human movements. This research advances our knowledge of human motion through video analysis and bridges the gap between technology and artistic expression.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Machine learning**; • **Applied computing**;

KEYWORDS

Human-AI Collaboration, Dance Movement Analysis, Machine Learning in Art, Video Segmentation

ACM Reference Format:
John Gunerli, Manoj Deshpande, and Brian Magerko. 2024. Video Segmentation Pipeline For Co-Creative AI Dance Application. In *9th International Conference on Movement and Computing (MOCO '24), May 30–June 02, 2024, Utrecht, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3658852.3659085

1 INTRODUCTION

The ability to accurately capture, analyze, and interpret human motion from video data is pivotal in enhancing our understanding of performance optimization, human movement analysis, and dance movement understanding. In the realm of dance and performance arts, motion capture and its analysis could potentially enhance understanding of dance movements. Utilizing computer vision (CV) for motion capture offers several advantages over utilizing motion capture suits for motion tracking. First and most importantly, it is non-invasive, enabling for the observation and recording of movements without physical interference or discomfort to artists, allowing them to freely and naturally express themselves. This aspect is critical in fields like dance, where the authenticity of movement is vital for body-action classification and movement analysis. Unlike motion tracking technologies that may require time-consuming setups, video capture is more straightforward, with minimal setup time, making it highly efficient for spontaneous and planned recording sessions. Furthermore, CV-based motion analysis eliminates the need for specialized proprietary software or hardware, often associated with traditional motion tracking systems. Traditional motion-capture systems, often reliant on suits equipped with sensors, face several challenges that can limit their effectiveness in capturing dance movements. Sensor noise can introduce errors into the data, especially in the dynamic and unpredictable environment of dance performances. [15] [18] Additionally, these systems typically capture only a limited number of key points on the dancer's body, which may not adequately represent the full complexity of dance movements that involve intricate gestures and subtle expressions. Relying on a CV-based approach for movement analysis not only makes it a cost-effective solution that avoids these limits mentioned above but also broadens its accessibility to a wider range of researchers and practitioners. It is also not uncommon to hear of accidents that occur where motion capture suits catch fire due to overheating of internal sensors or batteries, which puts dancers at risk. Additionally, there is no prerequisite for dancers to undergo specialized training to interact with the technology.

Interpreting human motion from videos, particularly in dance, is challenging due to the intricate and expressive nature of the movements. A basic CV algorithm typically falls short of extracting data of sufficient quality for training machine learning models. A more sophisticated strategy is necessary, which involves breaking down the overall task into smaller, manageable subtasks and employing a series of specialized algorithms in a coordinated sequence to create a computational pipeline. By doing so, we can effectively generate high-quality data suitable for further training and analysis. However, given the complexity of dance as an art form, it is crucial that the pipeline allows for human feedback at every subtask stage to ensure the accuracy and relevance of the analysis.

Our approach leverages computer vision and machine learning to create a novel video processing and human-in-the-loop pipeline. This pipeline is designed to not only address the technical challenges of motion capture but also to enrich our understanding of

MOCO '24, May 30–June 02, 2024, Utrecht, Netherlands

ACM ISBN 979-8-4007-0994-4/24/05
https://doi.org/10.1145/3658852.3659085

the nuanced dynamics of human movement. By utilizing video processing tools and computational techniques, we aim to bridge the gap between understanding complex human motion, its digital analysis, and interpretation. This is particularly crucial in dance, where every gesture and movement carries a depth of meaning and emotion. Furthermore, integrating human feedback in our methodology reflects a commitment to pushing the boundaries of how technology can complement and enhance human skill and creativity. The pipeline we have developed is a part of the co-creative AI dance application LuminAI and consists of four intricately designed processes that work together to transform raw video data into a rich, analyzable format with timestamps. The rest of the paper is structured as follows. We first describe the related work that is composed of three sections — video segmentation pipeline, Human-AI co-creative systems, and LuminAI — followed by the core components of the proposed pipeline i.e., interactive image segmentation, batching, keyframe extraction, and precise timestamping - to address complex challenges described in related work in video-based movement analysis.

2 RELATED WORK

2.1 Video Segmentation

Video segmentation is crucial in movement analysis as it separates the person from the background. Various algorithms for video and image segmentation have been explored in the literature. Ferreira et al. [2] used transformer architecture, a type of Neural Network, for video segmentation, highlighting their strength in temporal analysis which is vital for tracking movement changes over time. However, their model, tailored for workout routine analysis, struggles with complex dance movements. Puertas et al. [14] applied multi-task learning to understand body poses and movements from still images, but this method falls short in classifying or segmenting multi-limb movements essential for dance.

Kuang and Tie [4] developed a novel flow-based encoder-decoder network (FUNet) for detecting human head and shoulders in videos, aiding in applications like video conferencing and virtual reality. This method overcomes issues like motion blur from rapid head movements or hand waving but is limited to upper-body movements. These challenges in understanding, classifying, and segmenting body actions and gestures involving multiple body parts are further highlighted in Liu et al. [6]'s work.

Our proposed method aims to tackle these issues. We utilize two pre-trained models, MagicTouch and PoseLandmarker from Google's MediaPipe suite, based on a Convolutional Neural Network (CNN) architecture, which provide better results as highlighted in the FUNet paper. These already proven and integrative models are crucial for real-time assessment and interpretation of human gestures and movements. PoseLandmarker, unlike Puertas et al.'s approach, offers a more comprehensive understanding of body poses and movements. Unlike FUNet, these integrative models cover full-body movement analysis and not just the upper body. This combination of techniques significantly improves the ability to capture subtle nuances in body movements, which is essential for detailed dance analysis. We will discuss our method in more detail in the following section.

2.2 Human-AI Co-Creative Systems

Long et al. [10] emphasize the significance of AI in public spaces and its positive public perception. The progress in video segmentation and gesture classification goes beyond mere technical advancements in understanding human movements; it plays a vital role in preparing data for enhancing interactive experiences in Human-AI collaboration systems. Such pipelines, like the one we propose, enable Human-AI systems to grasp the intricacies and nuances in artistic expressions such as dance.

Developing Human-AI systems capable of collaborating in dance is challenging. LaViers [5] points out the necessity for these systems to mimic human movement's fluidity and subtlety rather than just focusing on precision and speed. This calls for a qualitative approach to classifying human movements, incorporating insights from performance arts. A taxonomic framework that blends qualitative and quantitative measures is essential to overcome these challenges. According to Zhu [19] , the ultimate aim for machine systems is not only to respond but also to adapt to human nuances, thereby enhancing their collaborative abilities. This requires developers to integrate data-driven and knowledge-driven approaches, promoting advanced learning and decision-making in Human-AI systems.

Such integration demands innovative architectures and models adept at processing complex inputs and facilitating intuitive, human-centric interactions. Hence, in our proposed pipeline, we strive to combine these sophisticated techniques and concepts. Our overall goal is to develop a system that captures and analyzes dance movements with high precision which is able to comprehend and adapt to the subtleties and nuances of human expression in dance.

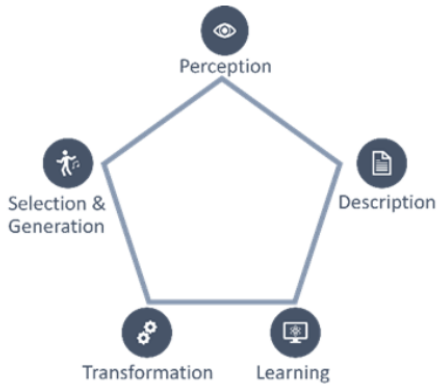

Figure 1: Five Different Software Stages of LuminAI.

2.3 LuminAI

LuminAI, is an interactive installation that features a virtual AI agent capable of improvising movements in collaboration with human participants [3, 11]. This system leverages computational models based on improvisational theories from theater and dance.

Over several years of development, LuminAI has generated diverse research outcomes in designing co-creative agents for public spaces such as museums [9]. Key areas of exploration include implementing machine learning techniques for real-time, embodied improvisation [7, 8] and understanding the influence of creativity heuristics (novelty, value, and unexpectedness) on the creative process of collaborative movement improvisation [10]. The research on LuminAI has contributed to our understanding of Human-AI interaction and has been foundational to the field of computational co-creativity.

The focus of LuminAI's current research has shifted from providing a casual, interactive dance experience to a more specialized study of the improvisational process in dyadic dances involving professional dancers. The aim is to redesign LuminAI so that it can improvise in real-time with trained and professional dancers, enhancing the depth and complexity of the interactive dance experience.[16]

Currently, LuminAI comprises five distinct modules, each with a specific function as shown in Figure-1. The first module is the perception module, which is responsible for recognizing and segmenting continuous motion capture data into individual body actions. The second, the description module, interprets these body actions in relation to each other and based on implicit domain knowledge. The third is the learning module, which assimilates patterns of body actions and movements. The fourth module, the collaboration dynamics module, analyzes sequences of body actions to discern the social dynamics of the interaction. Finally, the fifth module is the body action generation and selection module. This module creates potential dance movement responses using the database of common improvisational movements and then selects a response based on the characteristics of the body actions and the understood collaboration dynamics. The work in this paper pertains only to the perception module.

3 VIDEO SEGMENTATION PIPELINE

In this section, we will detail the proposed video processing pipeline. The pipeline consists of four separate sub-processes as shown in Figure-2

3.1 Interactive Image Segmentation

This part of the pipeline aims to track the location of the particular person in the frame of a video. This complex task is broken down into several critical sub-tasks. First, we have Pose Estimation, where the sub-process pinpoints key body landmarks (like shoulders and hips) to understand the orientation and movements of the person followed by Person Segmentation, where the part identifies and isolates the person in the video frame, distinguishing them from other elements such as the background or other objects using information from key body landmarks. Another sub-task is Frame-by-Frame Analysis, ensuring consistent tracking of the individual across each frame of the video, with continuous updates on their position and pose. This part of the task is crucial, as it ensures that if the person becomes undetectable by the computer vision algorithms at any point, the pipeline maintains continuous tracking and does not "lose" the person in question. This leads to the Segmentation phase, where the individual is segmented from the rest of the frame, allowing for focused analysis. Tackling these sub-tasks enables the pipeline to effectively focus on and analyze the presence of an individual and movements in the video, which is crucial for the rest of the pipeline to work effectively.

The Pose Detection model [12] [1] excels in breaking down video into individual frames and identifying key body landmarks available by MediaPipe such as the left shoulder, right shoulder, right hip, and left hip. However, it falls short of precisely locating the human subject within these frames and segmenting them out of the images. On the other hand, the Image Segmentation model [13][12] alone is adept at discerning the exact pixels where the human is present, enabling the segmentation of the person from each frame. However, without the initial pinpointing of the human subject, its efficiency is greatly hindered. In short, without pose landmarks, the image segmentation is incapable of locating the person in each individual frame (Figure 3), and without the image segmentation, the pose landmark estimation does not allow the systems to focus on the human's movements.

Thus, this process utilizes two specialized pre-trained models from MediaPipe: the Pose Detection model and the Interactive Image Segmentation model. The pose detection model from MediaPipe takes in the video and turns the video into individual frames. This process allows MediaPipe to recognize the person at each individual frame, conduct pose landmark detection to predict/identify four key body landmark locations i.e., Left Shoulder, Right shoulder, Right Hip and Left Hip. These points are then fed onto the interactive segmentation model, which uses this information to obtain the exact pixels in which the human is present, allowing us to segment the person out of each frame. The reason to use two models is to use the pose landmark model to obtain the precise location of the user, which then allows the interactive image segmentation to know exactly where the human is located in a particular frame (Figure 4).

3.2 Batching

Batching process is a human-in-the-loop system and the output is highly dependent on the choice of the user. The user gets to provide the segment durations and the frame rate of the original video by themselves with no default value provided since there are unique requirements that a standard solution may not be able to address effectively. The batching creates sub-clips of the interactively segmented video, to be processed by the keyframe extractor later. This step is particularly beneficial for processing large video files and facilitates parallel processing, enhancing overall efficiency. Such involvement is particularly beneficial for processing large video files, as it allows for customized handling and optimization of data. From our empirical observation in applying the batching process, setting shorter segment durations has proven effective in studies focusing on quick, intricate dance routines where capturing every nuance is crucial. It must be noted that the changes in segment durations also heavily affect processing time. For performances that involve slower, more expressive movements, longer segment durations have facilitated a more manageable volume of data without compromising the quality of motion capture.

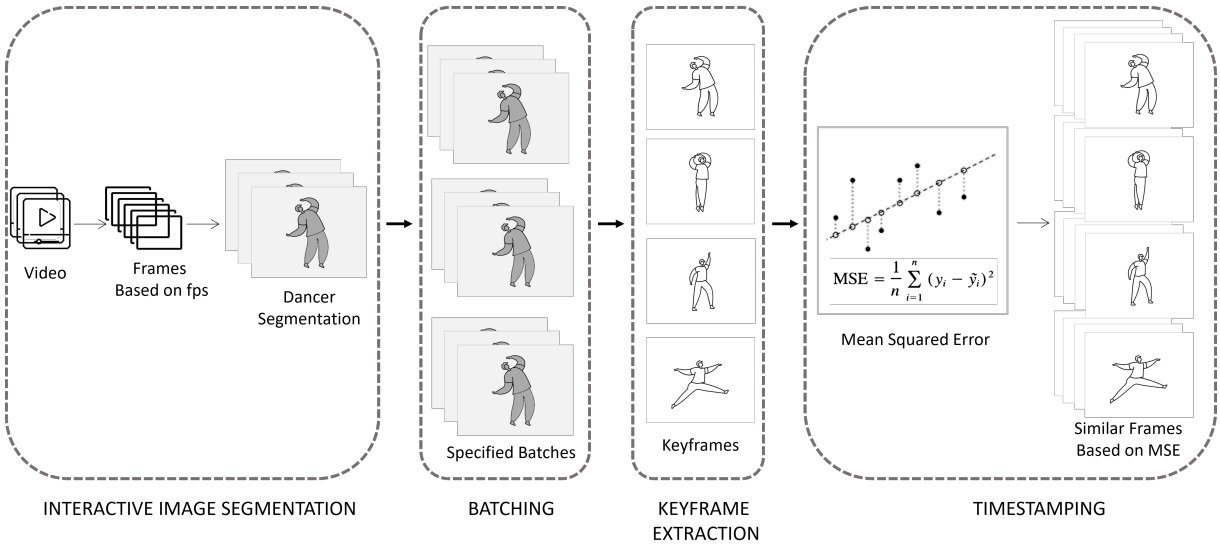

Figure 2: Video Segmentation Pipeline.

Figure 3: Image segmentation only

Figure 4: Pose detection + Interactive image segmentation

3.3 Keyframe Extraction

Following the batching process that creates the sub-clips, the keyframe extraction is done using an open-source module called Katna. Katna takes in a video and selects a designated number of frames that represent the video content of the batched video. To select the appropriate keyframe, Katna utilizes several selection criteria to ensure the extracted frames provide an accurate and compact summary of the video content.

- K-Means Clustering is employed on the frames, utilizing their image histogram to group similar frames together. This technique aids in organizing the video content and simplifying the selection process. The K within this context is determined by the number of keyframes the user desires across the total number of frames.
- Absolute Differences in the LUV Color Space: Within each cluster, frames are selected based on absolute differences in the LUV color space to assess the uniqueness of each frame. This method ensures that each chosen frame represents a significant visual change, capturing key moments that are distinct within their clusters.
- Brightness Score Filtering: This step applies filtering to ensure that the selected frames have optimal lighting, which enhances the visual quality and clarity of the keyframes.
- Variance of the Laplacian: The best frame from each cluster is selected based on the variance of the Laplacian, which measures image sharpness. This criterion ensures the selection of the sharpest and most representative frame from each group.
- Entropy and Contrast Score Filtering: Frames are further evaluated for their clarity and detail based on entropy and contrast scores. Frames with higher scores are preferred as they provide a clearer and more detailed summary of the video content.

3.4 Timestamping

The next phase focuses on accurately time-stamping frames within the batched video, by integrating numerical computation libraries and video/image editing tools such as Matplotlib, NumPy and MoviePy, Pillow respectively. The method applies a mean squared error (MSE) calculation to precisely find timestamps of where the

keyframe was extracted from the batched video frames. This technique ensures high precision in identifying the temporal characteristics of each frame, crucial for chronological analyses. These timestamps are not merely chronological markers but serve as a vital reference for correlating the keyframes with specific moments in the original video. The reason as to why MSE calculation is preferred over other SOTA techniques such as cosine similarity and image embeddings is because of the fact that MSE provides a more direct measure of the pixel-by-pixel differences between frames. This granular approach is crucial for our application, where subtle variances in movement and positioning can have significant implications for the analysis. It must also be noted that image embeddings and cosine similarity are also effective ways to accomplish these tasks, but they come with a high level of computational complexity, as well as time complexity for large frame comparison tasks, and they may not offer the same level of precision in detecting minute differences between frames as required in our study.

4 DISCUSSION AND FUTURE WORK

Integrating a human-in-the-loop system into our video processing pipeline is a significant step forward in human motion analysis, especially for dance and artistic expression. By incorporating human input at crucial stages, we ensure the system aligns intuitively with the nuanced demands of analyzing human movement. Our pipeline tackles the challenge of segmenting complex dance movements involving multiple body parts by utilizing multiple specialized algorithms for specific sub-tasks. This approach allows for tailoring and fine-tuning the process according to the unique characteristics of the video data and the specific artistic or analytical objectives of each project. The human-centric nature of the pipeline ensures that the technology goes beyond being a simplistic, one-click black-box solution, and instead is continuously shaped and enhanced by human experience and perception.

Currently, the Video Segmentation Pipeline only works with a single person in the frame to simplify the process of keyframe extraction and batching processes. In the next iteration, analysis of movement where multiple people are present can greatly allow for the pipeline to achieve a wide variety of applications like choreography analysis and enhancement in dance performances for understanding the importance of movement in collaborative movement analysis. It must also be noted that the process of classifying and analyzing dance body actions is complex and multifaceted. After segmenting videos, it is essential to identify and classify body actions. Movement analysis frameworks such as the Laban Movement Analysis (LMA) [17] are crucial to help us understand the categorization and quality of movements.

In this paper, we demonstrated the development of a human-in-the-loop video segmentation pipeline, tailored specifically for the intricate analysis of human motion in the context of dance. The research enhances our technical understanding of human movement via video analysis and contributes to the blend of technology and the arts. The pipeline's core components - interactive image segmentation, batching, keyframe extraction, and precise timestamping - work harmoniously to address complex challenges in video-based movement analysis. This approach not only enhances the accuracy of motion capture and interpretation, but also respects the intricacy and subtlety inherent in human movement, particularly in the art of dance.

ACKNOWLEDGMENTS

This work was supported by NSF Grant #2123597.

REFERENCES

[1] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device Real-time Body Pose tracking. arXiv:2006.10204 [cs.CV]

[2] Bruno Ferreira, Paulo Menezes, and Jorge Batista. 2022. Transformers for Workout Video Segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*. 3470–3474. https://doi.org/10.1109/ICIP46576.2022.9897194

[3] Mikhail Jacob and Brian Magerko. 2015. Viewpoints AI. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition* (Glasgow, United Kingdom) *(C&C '15)*. Association for Computing Machinery, New York, NY, USA, 361–362. https://doi.org/10.1145/2757226.2757400

[4] Zijian Kuang and Xinran Tie. 2021. Flow-based Video Segmentation for Human Head and Shoulders. arXiv:2104.09752 [cs.CV]

[5] A. LaViers. 2019. Make robot motions natural. *Robotics* (2019).

[6] L. Liu, D. Long, S. Gujrania, and B. Magerko. 2019. Learning movement through human-computer co-creative improvisation. In *Proceedings of the 6th International Conference on Movement and Computing*.

[7] Lucas Liu, Duri Long, Swar Gujrania, and Brian Magerko. 2019. Learning Movement through Human-Computer Co-Creative Improvisation. In *Proceedings of the 6th International Conference on Movement and Computing* (Tempe, AZ, USA) *(MOCO '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 8 pages. https://doi.org/10.1145/3347122.3347127

[8] Lucas Liu, Duri Long, and Brian Magerko. 2020. MoViz: A Visualization Tool for Comparing Motion Capture Data Clustering Algorithms. In *Proceedings of the 7th International Conference on Movement and Computing* (Jersey City/Virtual, NJ, USA) *(MOCO '20)*. Association for Computing Machinery, New York, NY, USA, Article 9, 8 pages. https://doi.org/10.1145/3401956.3404228

[9] Duri Long, Mikhail Jacob, Nicholas Davis, and Brian Magerko. 2017. Designing for Socially Interactive Systems. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition* (Singapore, Singapore) *(C&C '17)*. Association for Computing Machinery, New York, NY, USA, 39–50. https://doi.org/10.1145/3059454.3059479

[10] D. Long, M. Jacob, and B. Magerko. 2019. Designing co-creative AI for public spaces. In *Proceedings of the 6th International Conference on Movement and Computing*.

[11] Duri Long, Lucas Liu, Swar Gujrania, Cassandra Naomi, and Brian Magerko. 2020. Visualizing Improvisation in LuminAI, an AI Partner for Co-Creative Dance. In *Proceedings of the 7th International Conference on Movement and Computing* (Jersey City/Virtual, NJ, USA) *(MOCO '20)*. Association for Computing Machinery, New York, NY, USA, Article 39, 2 pages. https://doi.org/10.1145/3401956.3404258

[12] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. arXiv:1906.08172 [cs.DC]

[13] MediaPipe. 2023. Interactive image segmentation task guide | MediaPipe. https://developers.google.com/mediapipe/solutions/vision/interactive_segmenter

[14] E. Puertas, M. A. Bautista, D. Sanchez, S. Escalera, and O. Pujol. 2015. Learning to Segment Humans by Stacking Their Body Parts. In *Computer Vision - ECCV 2014 Workshops*, Lourdes Agapito, Michael M. Bronstein, and Carsten Rother (Eds.). Springer International Publishing, Cham, 685–697.

[15] Ståle Andreas van Dorp Skogstad. 2014. *Methods and Technologies for Using Body Motion for Real-Time Musical Interaction.* Ph. D. Dissertation.

[16] Milka Trajkova, Manoj Deshpande, Andrea Knowlton, Cassandra Monden, Duri Long, and Brian Magerko. 2023. AI Meets Holographic Pepper's Ghost: A Co-Creative Public Dance Experience. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 274–278. https://doi.org/10.1145/3563703.3596658

[17] R. von Laban and L. Ullmann. 1971. *The Mastery of Movement.* Macdonald & Evans. https://books.google.com/books?id=-RYIAQAAMAAJ

[18] Graeme A. Wood. 1982. DATA SMOOTHING AND DIFFERENTIATION PROCEDURES IN BIOMECHANICS. *Exercise and Sport Sciences Reviews* 10, 1 (1982).

[19] S. Zhu, T. Yu, T. Xu, H. Chen, S. Dustdar, S. Gigan, D. Gunduz, E. Hossain, Y. Jin, F. Lin, B. Liu, Z. Wan, J. Zhang, Z. Zhao, W. Zhu, Z. Chen, T. Durrani, H. Wang, J. Wu, and Y. Pan. 2022. Intelligent computing: The latest advances, challenges and future. *arXiv preprint arXiv:2211.11281* (2022).