A Real-time Pose Estimation Application for Tinikling

A Capstone Project on Operational Technologies
Presented to the Faculty of the
Department of Electronics and Computer Engineering
Gokongwei College of Engineering
De La Salle University

In Partial Fulfillment of the
Operational Technologies

by

CALAGUIAN Nathan Raekel L.
ELLAR Gerald Antonio P.
MAHAIT Hans

October, 2025

ABSTRACT

Index Terms—Dance, Pose Estimation, Real-time, OpenPose .

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

NOTATION

100 # GLOSSARY

101	Tinikling	The traditional Filipino dance involving two bamboo sticks, where a dancer moves in and out of the rhythmically tapped sticks.
102	OpenCV	An open-source computer vision library widely used for real-time image capture and processing, including camera I/O, preprocessing, filtering, and contour extraction.
103	Ultraleap	A commercial infrared-based hand-tracking system that uses stereo cameras and near-infrared illumination to generate dense, low-latency 3D hand data.
104	Hand Gesture Recognition	The computational pipeline of detecting a hand, extracting features or landmarks, and classifying the resulting input into a specific gesture with low latency and high reliability.
105	MediaPipe	A framework for building multimodal applied machine learning pipelines, including computer vision models like hand gesture recognition.
106	Pose estimation	A computer vision technique used to detect human poses (such as hand or body positions) from images or videos, often used for gesture and movement analysis.
107	Dynamic Time Warping	A technique used to align time sequences by minimizing the distance between them, useful for applications like gesture recognition and speech processing.
108	Operational Technologies	Programmable systems or devices that interact with the physical environment (or manage devices that interact with the physical environment). These systems/devices detect or cause a direct change through the monitoring and/or control of devices, processes, and events. Examples include industrial control systems, building management systems, fire control systems, and physical access control mechanisms.

109

LISTINGS

Chapter 1

INTRODUCTION

1.1 Background of the Study

Classical Computer Vision (CV) approaches used skin color segmentation, contour analysis, optical flow, and handcrafted descriptors (Histogram of Oriented Gradients (HOG), motion history images) to detect and classify gestures. Despite being simple and interpretable, those methods struggle with background variation and scale. The deep-learning era replaced handcrafted features with Convolutional Neural Network (CNN) that learn hierarchical visual features directly from image data, yielding much higher accuracy for static hand pose and short-sequence recognition tasks. Many recent capstone and journal implementations pair OpenCV (for capture/preprocessing) with CNN built and trained in TensorFlow/PyTorch to recognize a fixed vocabulary of gestures in real time. These hybrid pipelines are practical for capstone projects because OpenCV handles efficient frame processing while CNNs provide generalization across users and backgrounds. Furthermore, Operational Technologies plays a crucial role in deploying these systems in real-world applications where physical devices and processes are monitored and controlled, such as in industrial automation or building management systems, which benefit from enhanced gesture recognition. (Oudah et al., 2020)

Instead of classifying raw images, several high-performance systems first extract skeletal landmarks (e.g., MediaPipe's 21-point hand model) and feed those coordinates to a classifier (small CNN, MLP, or temporal model like LSTM). Landmark-based pipelines reduce sensitivity to background and scale and make models smaller and faster, which is ideal for mobile or AR deployment. Markerless commercial devices such as the Leap Motion Controller and Ultraleap cameras provide very accurate 3D joint data using IR illumination and multi-camera setups; those give superior fidelity but add hardware cost and integration work. For a capstone aiming at broad deployability, a practical approach is to prototype with MediaPipe + OpenCV + CNN (or lightweight temporal model) and consider Ultraleap integration later for high-precision installations. /citepzhangmediapipe

1.2 Prior Studies

MediaPipe Hands presents a two-stage on-device pipeline (Zhang et al., 2020)(palm detector + hand-landmark regressor) that extracts 21 hand landmarks from a single RGB frame and runs in real time on mobile GPUs; the architecture and open implementation are widely used as a practical basis for gesture recognition because they offer compact, robust landmark outputs that are easier to classify than raw images. This work is especially relevant to mobile or cross-platform deployment without extra hardware.

Reviews and vendor docs show that Ultraleap's IR stereo cameras and LED illumination give very precise 3D joint tracking and low latency (Ultraleap, 2025), making them popular for VR/installation work; academic comparisons find Leap/Ultraleap and MediaPipe are both capable, with trade-offs in precision versus hardware requirements. Ultraleap or similar IR camera hardware is a practical choice for professional installation quality (amusement park kiosk, VR attraction). American Sign Language (ASL) and other sign recognition papers demonstrate that combining landmark features (from MediaPipe or depth sensors) with temporal models (Long Short-Term Memory (LSTM)/CNN temporal stacks) yields state-of-the-art results for complex hand sequences (Verma et al., 2024). These studies emphasize the importance of considering variable visibility conditions as spellcasting often requires temporal tracing (drawing shapes), and not just static poses. This also provides insight into dataset design and labeling strategies. Hand Gesture Recognition is advancing as a key technology for human–machine interaction. This study reviews both non-vision (e.g., sensor-based) and vision-based approaches, examining tools such as hidden Markov models (Rahman et al., 2025), finite state machines, color modeling, naive Bayes, deep networks, histogram features, and fuzzy clustering. Methods are categorized into detection, tracking, and recognition phases, with comparisons across static and dynamic gestures. The review highlights current technologies, their advantages and limitations, and identifies directions for future research.

Hand gestures, as a form of nonverbal communication, are applied in fields such as HCI, assistive communication, robotics, home automation, and healthcare (Oudah et al., 2020). Research spans sensor-based and vision-based methods, with gestures categorized as static, dynamic, or hybrid. This paper reviews literature on gesture recognition, comparing techniques in terms of segmentation, classification, datasets, gesture types, camera use, detection range, and performance. It provides a comprehensive overview of methods, their merits and limitations, and potential applications.

1.3 Problem Statement

Immersive interactive systems in gaming, AR, amusement parks, and accessibility still rely heavily on handheld controllers, touchscreens, or specialized hardware that break immersion, add cost, or exclude users with differing motor abilities. Markerless, camera-based hand-gesture recognition promises touchless, expressive input suitable for "magical" metaphors (casting spells, tracing runes) that are intuitive and socially engaging. However, real-world deployment is challenged by variable lighting, occlusion, noisy backgrounds, and latency. These problems make accuracy and robustness the central obstacles for any spellcasting CV system. Modern solutions that combine real-time hand-landmark extraction and convolutional neural networksCNN have narrowed the gap, but careful design is required to meet the high level competency goals for responsiveness, cross-platform deployment, and accessibility. (Oudah et al., 2020)

1. **PS1:**

    - The ideal scenario for our intended audience (students, educators, and dance enthusiasts) is to have an intuitive and interactive learning tool that facilitates the practice of Tinikling, the traditional Filipino dance. This tool should provide on users' dance movements, enabling them to learn and improve their technique. The desired state includes accessibility to the tool on various devices (e.g., desktop, mobile) with a user-friendly interface and a high level of accuracy in tracking the dance steps. Additionally, it should support personalized feedback, enabling users of all skill levels to progress and feel engaged in learning this cultural heritage.

2. **PS2:**

    - Currently, learning Tinikling requires access to physical dance classes or instructors, which are often limited by geographical location, financial resources, or time constraints. For individuals unable to attend such classes, the lack of affordable and effective learning tools becomes a significant barrier. Additionally, existing dance-learning technologies are either costly, relying on specialized hardware, or lack the immediacy of real-time feedback, making it difficult for learners to practice and perfect their movements without direct instructor guidance.

- The pain point is that students who want to practice Tinikling at home or in remote areas are unable to receive real-time guidance or feedback, leading to slower progress, incorrect technique, and a loss of motivation.

3. **PS3:**

- Without a tool that offers immediate feedback and a clear learning path, students practicing Tinikling on their own are likely to struggle with incorrect movements, which may lead to frustration. Over time, this lack of progress could result in a lack of confidence, disengagement from the learning process, and ultimately, the inability to learn the dance correctly. Furthermore, the absence of accessible learning tools risks the loss of cultural knowledge and the fading of the Tinikling tradition, especially among younger generations who may not have easy access to traditional learning methods.

1.4 Objectives and Deliverables

1.4.1 General Objective (GO)

- GO: To design and implement a real-time Pose estimation-based tinikling learning application that provides immediate feedback and scoring to users based on their performance of tinikling dance routines;

1.4.2 Specific Objectives (SOs)

- SO1: To implement a real-time pipeline that captures camera frames, extracts robust hand features (landmarks or processed images), and classifies gestures into a configurable spell vocabulary with low latency (30 fps target) and high accuracy;;
- SO2: To make the model robust to lighting, background clutter, and user variation through and landmark-based representations ;
- SO3: To design the system to be deployable across desktop, mobile, and simple AR setups using cross-platform libraries (OpenCV, MediaPipe, TensorFlow/TensorFlow Lite) ;
- SO4: To make the interaction ergonomically accessible by supporting alternative gestures and calibration for users with different ranges of motion ;
- SO5: On UX side, to make spells feel immediately meaningful (clear mapping between motion and effect), provide instant feedback when a spell is recognized, and allow easy extension of the spell set. ;

1.4.3 Expected Deliverables

1.5 Significance of the Study

This capstone project focuses on the development of a tinikling learning application through the integration of pose estimation and human action recognition. The setup consists of a webcam, laptop, and two bamboo sticks for the tinikling dance. Such a setup offers affordability and accessibility benefits for users. Ultimately, it contributes to the field

TABLE 1.1 EXPECTED DELIVERABLES PER OBJECTIVE

Objectives	Expected Deliverables
GO: To develop a real-time pose estimation-based Tinikling learning application	= • Prototype of Tinikling learning application. • Documentation and user manual.
SO1: To develop a real-time pose estimation pipeline that captures the movement of dancers through a webcam, detects skeletal keypoints, and analyzes poses for Tinikling steps with low latency and high accuracy.	• Optimized skeletal keypoints detection for Tinikling steps. • Implementation of webcam-based pose estimation pipeline. • Performance evaluation results.
SO2: To make the pose estimation model robust to lighting, background clutter, and user variation through dataset collection, augmentation, and landmark-based representations.	• Augmented dataset covering varied lighting, backgrounds, and user types. • Enhanced landmark-based model with robustness improvements. • Comparative performance evaluation report.
SO3: To design and integrate a scoring and feedback system that evaluates users' dance accuracy in a post-performance review by aligning user poses with reference choreographies.	• Scoring and feedback algorithm. • Tinikling choreography database. • Post-performance scoring output with accuracy metrics.
SO4: To evaluate the system's performance and usability through controlled testing with dancers or students, measuring accuracy, latency, and user experience for future refinement and educational deployment.	• Conducted controlled testing with participants. • Collected performance and usability metrics. • Evaluation report with recommendations for improvement.

of both pose estimation and human action recognition by demonstrating a successful integration of the two in a live setup.

1.5.1 Technical Benefit

1. Enables real-time pose estimation and post-performance feedback, improving accuracy and efficiency throughout the learning process.

2. Low-cost software-based learning tool which uses a webcam and desktop computer rather than expensive motion capture equipment.

1.5.2 Social Impact

Promotes cultural preservation by making Tinikling more accessible through interactive applications.

Student engagement and participation in cultural education through gamifying the learning experience.

Supports remote or in-classroom instruction through allowing instructors to integrate technology with dance education.

1.5.3 Environmental Welfare

Utilizes existing and widely available hardware such as webcams and desktop new specialized equipment, which ultimately lessens electronic waste.

Encourages digital preservation of cultural heritage, lessening reliance on physical materials be it through either physical archives or infrastructure.

1.6 Assumptions, Scope, and Delimitations

1.6.1 Assumptions

1. Pose landmarks from consumer-grade RGB cameras or low-cost depth sensors provide sufficient fidelity to represent tinikling movements for temporal alignment

and scoring.

2. Choreography can be divided into short, labeled segments that enable reliable matching and targeted feedback.
3. Dynamic Time Warping or a constrained variant will handle tempo variation robustly for temporal alignment.
4. A brief per-user calibration step will improve scoring consistency.

1.6.2 Scope

1. Cover automatic pose estimation, sequence alignment, and segment-level scoring for tinikling.
2. Accept landmark or depth inputs and provide immediate on-device cues during performance.
3. Produce a higher-precision final score after a more detailed pass.
4. Use self-sourced tinikling videos for model training when no public dataset exists.
5. Benchmark against general dance datasets where appropriate.
6. Report sensor-based metrics and simple user measures such as perceived accuracy and engagement.

1.6.3 Delimitations

1. Will not perform detailed facial or hand mesh reconstruction.

2. Will not replace multi-camera motion capture for research-grade kinematics.
3. Will not guarantee reliable results under heavy occlusion, very low light, extreme off-axis views, or when clothing blends with the background.
4. Will not attempt full generalization to all body shapes without additional data and tuning.
5. Limits reflect known sensor and algorithm constraints and the aim to produce a practical, lightweight prototype.

1.7 Description and Methodology of the Capstone Project on Operational Technologies

1. Phase 1: Model Development serves a precursor for Phase 2 wherein the specifics of the model, libraries, and environment to use are defined. In total, Phase 1 would last 4 weeks spanning from week 4 to 7. The bulk of the research for the project would be carried out during this phase.The dataset to be used for training would be collected during this phase as well.
2. Phase 2: Model Training consists of training the model using the dataset collected in the previous phase. This phase will largely consist of testing and improving the resulting model. Tests would be conducted using the group members as dancers. This phase also includes the optimization of the model for real-time detection simultaneously with the music. In total, this phase would last 4 weeks spanning from week 8 to 11.

De La Salle University

3. Phase 3: UI/EX Development consists of the integration of the trained model with a user interface. Once integrated final testing and refinement of the final program would be carried out. The final output would be presented as well during this phase along with the finalization of the documentation. This phase would last for 3 weeks spanning from week 11 to 13

1.8 Estimated Work Schedule and Budget

1.8.1 Milestones and Gantt Chart

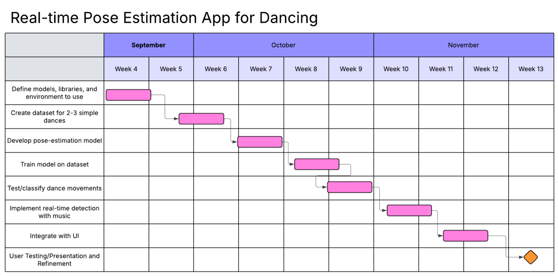

Fig. 1.1 Milestone Gantt Chart for Real-time Pose Estimation Dance Software

1.8.2 Budget

Given that the capstone project largely consists of software, apart from the use of a laptop for both programming, as well as actual implementation and usage of the dance program, the only expense to consider would be for that of a Webcam, which is already owned.

TABLE 1.2 OPERATIONAL FINANCIAL PLAN

Item	**Price**
Webcam	₱1,850
Item	**Price**
4pc. Tinikling Sticks	₱110
Total	**₱1,960**

1.9 Overview of the Capstone Project on Operational Technologies

This capstone project focuses on developing a real-time pose estimation-based learning application for Tinikling, the Philippine national dance. It integrates computer vision and machine learning techniques in order to create an interactive learning platform that performance scoring to users. The project utilizes of webcams, MediaPipe-based skeletal landmark extraction to analyze users' movements relative to reference choreography. Unlike expensive motion capture systems, this setup uses low-cost and accessible hardware, making the system practical for classroom, cultural, and home use. The system emphasizes cultural preservation by modernizing Tinikling education through technology. It enables students to learn and practice the dance interactively, provides technical benefits such as real-time feedback without costly sensors, and supports social and environmental goals through cultural engagement and sustainable use of existing hardware.

Chapter 2

LITERATURE REVIEW

2.1 Existing Work

A study by Venkatrayappa et al. (2024) focused on surveying the various existing 3D human body pose and shape estimation techniques, given its crucial nature in fields such as augmented or virtual reality, healthcare and fitness technology, and virtual retail. The solutions explored consisted of mainly three types of inputs, which were single images, multi-view images, and videos. Various issues pertaining to dance, such as fast motion, occlusion, and unusual poses were analyzed to see how each affected the performance of each method. The specific models consisted of SMPL -A, SMPL -X, MANO, STAR, FLAME, which are optimization-based models, as well as HMR, VIBE, SPIN, PARE, EXPOSE, and PHALP, which are deep learning-based models. SMPL was found to be beneficial in terms of realistic body representation, efficiency for real time applications, and wide availability, however it has limitations in areas pertaining to facial and hand modeling, as well as representation of ethnic diversity. SMPL -X proved to provide several advantages such as facial expressions, hand gestures, and improved expressiveness. Its limitations, however, consisted of simplified hand modeling and its limited pose variability. MANO offers detailed hand gesture modeling and realistic hand deformations, but has limitations due to its focus being exclusively on the modeling of hands, as well as computational challenges. STAR leverages sparse coding and temporal modeling, which allowed for a much more powerful framework for pose estimation., depicting state-of-the-art results throughout various benchmarks and practical implementations in sports analysis, human-computer interaction, and VR. FLAME was advantageous when it comes to computational efficiency, which made it suitable for real-time applications of pose estimation. As for its limitations, it primarily focuses on facial and lip modeling, which introduces complexity and potential computational challenges. MANO. HMR produces richer and more useful mesh representation, which is parameterized by shape and 3D joint angles. The network implicitly learns the angle limits of each joint. As such its use is discouraged for people with unusual body shapes. Its re-projection loss is highly under-constrained and it needs adversarial supervision in order to avoid unrealistic outputs. VIBE makes use of CNNs, RNNs and GANs, as well as a self-attention layer in order to achieve state-of-the-art results. A motion discriminator is used to help produce more realistic motion. Ultimately, the model is a standard SMPL body model format with sequences of poses and shape parameters. SPIN makes use of a self improving loop wherein better fits allow the network to train in a much more efficient manner while b.etter initial estimates from the network aids the optimization routine in order to result in better fits. PARE consists of a guided attention mechanism which exploits information on visibility of individual body parts all the while leveraging information from neighboring body parts in order to predict parts which are occluded. EXPOSE includes body, face, and hand estimation. It is able to estimate expressive 3D humans in a much more accurate manner in comparison to existing optimization methods at only a fraction of the computational costs. PHALP out performes all of the aforementioned methods. Despite this, it still has its limitations as well such as its reliance on a single camera, which may lead to issues such as occlusion and motion blur. It may also not work well in low-light conditions or when a person's clothes is a similar color to that of the background. Lastly, it also requires a significant amount of computational resources, which may make it not suitable for real-time applications.

A study by Protopapadakis et al. (2018), analyzes the effectiveness of various classification techniques in recognizing different dance types based on motion-capture skeleton data. Classifiers explored consisted of k-Nearest Neighbors (k-NN), Naïve Bayes, Discriminant

Analysis, Classification Trees, Random Forests (TreeBagger), Support Vector Machines (SVMs), and Ensemble Classifiers. Poses are identified through the use of body joints via Kinect sensor. The data set used consisted of various dances such as Enteka, Kalamatianos, Syrtos (Two-beat), Sytros (Three-beat). The kinect was used to capture skeletal joining data, to which feature extraction techniques such as principal component analysis and frame differencing were used in order to improve the classification accuracy. Ultimately, results showed that k-nearest neighbors and random forests are the best-performing classifiers among those that were explored. It was also proposed that the use of mulit-sensor or multimodal data may serve as a potential solution for issues specific to pose recognition in dance such as occlusion and complex movement patterns.

A study by ZZhao et al. (2025), looks into dance pose estimation and introduces the model DanceFormer. DanceFormer is a transformer-based model for dance pose estimation which makes use of the Vision Transformer, Time Series Transformer, and an edge computation layer in order to achieve a deep fusion of multimodal features and to overall increase its accuracy and real-time performance. The AIST and DanceTrack datasets were used throughout the experimentation. Results showed that DanceFormer out performs other models, with it achieving a pose estimation accuracy or MPJPE of 18.4mm and 20.1mm, as well as a multi-object tracking accuracy or MOTA of 92.3% and 89.5%. It is also suitable for real-time processing in even low-resource with an average latency of 35.2ms. Ultimately, it serves as an efficient, precise and real time solution for rather complex dance scenarios. It also has applications in a much more broad sense be it in dance education or in real-time motion analysis.

A study by Lei et al. (2023) discusses dance movement recognition based on gesture. A low accuracy traditional dance movement recognition algorithm based on human posture estimation was proposed. PAFs algorithm was used in order to recognize the spatial skeleton nodes and connections of joints in the human body. The pose of the body is estimated based on the movement of the spatial skeleton. Once the information on the detected posture is preprocessed and its features are extracted, LTSM time series algorithm was used in order to classify and recognize certain dance movements. Ultimately, results showed that the proposed algorithm has the capacity to reliably identify dance movements based on the skeleton nodes. It was able to achieve a recognition accuracy and recall rate upwards of 85% for the different movement categories. As for its recognition accuracy of curtsey movement, it achieved upwards of 95.2%.

Tölgyessy et al. (2021) present a detailed evaluation of Kinect v1, Kinect v2, and Azure Kinect skeleton tracking, analyzing joint-level error distributions and repeatability across distances and orientations. Their results highlight degradation in accuracy under occlusion, off-axis angles, and larger working distances, conditions typical of casual living-room dance setups. The findings underline both the potential and the limits of Kinect-class sensors, suggesting that practical applications often require either sensor fusion and smoothing to handle jitter or a focus on more reliable joints for robust real-time scoring.

Lin (2015) investigate how interactive feedback design influences user motivation in the context of Just Dance. Their study demonstrates that timely, clear cues significantly improve engagement, perceived competence, and sustained participation, with direct effects on physical activity outcomes. These findings show that feedback modalities and latency are as critical as recognition accuracy in shaping the player experience, emphasizing the importance of immediate, multimodal responses in dance or pose-based teaching applications.

Yu and Xiong (2019) propose and validate a Dynamic Time Warping method for evaluating rehabilitation exercises tracked with Kinect. Their algorithm successfully aligns noisy, tempo-varying motion with reference trajectories, producing reliable correctness scores even with partial occlusion. Applied to dance or short choreographies, DTW offers a robust foundation for handling tempo shifts and timing variation, supporting sequence-based scoring that is more forgiving than strict frame-to-frame comparison.

Rallis et al. (2019) compare Kinect II with the high-precision Vicon system in the context of choreography retrieval and analysis, using trajectory similarity measures such as DTW. While Kinect data contain noise and smoothing artifacts, the study shows that trajectory-level patterns remain useful when algorithms are designed to tolerate sensor bias. Their results support the use of low-cost consumer sensors, including RGB landmark pipelines, in applications where robust temporal alignment and trajectory modeling can offset hardware limitations.

Human pose estimation (HPE) has become an important area of study due to its applications in action recognition, sports, and performing arts. Xu, Zou, and Lin (2022) introduced the Adaptive Hypergraph Neural Network (AD-HNN), which captures high-order semantic dependencies among joints to improve multi-person pose estimation, particularly in handling occlusion and pose variability. In dance analysis, Ju (2025) applied deep learning with ResNet-152 and HR-Net to enhance dance pose recognition, addressing class imbalance and improving classification accuracy through global–local feature fusion.

For cultural preservation, motion capture (MoCap) has been widely adopted. Rizhan et al. (2025) demonstrated the use of MoCap to develop authentic motion templates for Malay folk dances, ensuring accuracy and authenticity in preserving intangible cultural heritage. In addition, Büyükgökoğlan and Uğuz (2025) developed a performance evaluation system for Turkish folk dances using deep learning–based pose estimation (e.g., Mediapipe, YOLO, LSTM), enabling objective assessment compared to traditional jury scoring.

TABLE 2.1 SUMMARY OF REVIEWED DANCE POSE ESTIMATION AND RECOGNITION STUDIES

Paper	Focus	Methodology	Results
Venkatrayappa et al. (2024)	Evaluates 3D human body pose & shape estimation methods for contemporary dance	Comparative survey of model families: SMPL(-A/X), MANO, STAR, FLAME (optimization-based) and HMR, VIBE, SPIN, PARE, EXPOSE, PHALP (learning-based); analysis by input modality (single-image, multi-view, video)	PHALP strong overall; SMPL-X improves expressiveness; STAR excels in temporal modeling; common limits are occlusion, lighting, and compute needs.
Protopapadakis et al. (2018)	Identifies dance types from motion-capture skeletal data	Kinect skeletal features with PCA + frame differencing; compared classifiers (k-NN, Naïve Bayes, LDA/DA, decision trees, Random Forest, SVM, ensembles)	k-NN and Random Forest performed best; multimodal data recommended to handle occlusion.
Zhao et al. (2025)	Real-time pose estimation for complex dances	Hybrid architecture: Vision Transformer + Time-Series Transformer trained on AIST and DanceTrack datasets	MPJPE: 18.4 mm / 20.1 mm; MOTA: 92.3% / 89.5%; latency 35.2 ms (real-time capable).

Continued on next page

Table 2.1 (continued)

Paper	Focus	Methodology	Results
Lei et al. (2023)	Improves recognition accuracy for traditional dance movements	Keypoint detection via Part Affinity Fields (PAFs); temporal modeling with LSTM classifiers	>85% overall accuracy; 95.2% for curtsey movements.
Zheng et al. (2023)	Deep-learning approaches for pose design and recognition	Backbone fusion (ResNet-152 + HR-Net) with global–local feature fusion and class-imbalance handling strategies	Reported metrics: accuracy 0.9870; precision 0.9851; sensitivity 0.9873; F1 0.9861; Kappa 0.9841.
Xu et al. (2022)	Multi-person pose estimation from single images	Two-stage Adaptive Hypergraph Neural Network (keypoint localization + adaptive hypergraph) with SIC module; end-to-end training	Achieves state-of-the-art performance on MS-COCO, MPII, and CrowdPose benchmarks.
Tölgyessy et al. (2021)	Quantifies joint-level accuracy and repeatability across Kinect sensors	Controlled robotic-manipulator and figurine measurements across positions; compared Kinect v1, v2, Azure Kinect (NFOV/WFOV)	Azure NFOV shows highest accuracy (0.8–1.9 mm SD); joint failures 15–30% under occlusion; performance declines at long ranges.
Lin (2015)	Effects of feedback and controller use in dance exergames	2×2×2 factorial experimental design (feedback × controller × sex); 129 participants; 12-minute sessions	Mean HR 109 bpm; immediate/clear feedback increased engagement and perceived competence.

Continued on next page

Table 2.1 (continued)

Paper	**Focus**	**Methodology**	**Results**
Yu and Xiong (2019)	DTW-based scoring for rehabilitation/exercise movements	Dynamic Time Warping on 8 bone vectors + body orientation; algorithm converting DTW distance to a 0–100% performance score	Scores correlated strongly with expert ratings (r 0.86); robust to tempo variation and some occlusion.
Rallis et al. (2019)	Choreographic pattern analysis from heterogeneous capture systems	Trajectory-based DTW similarity for choreography; sensor comparison (VICON vs Kinect); experiments on smoothing and joint selection	Kinect is noisier but DTW reduces sensor bias; smoothing and selective joint use improve retrieval accuracy.
Sun and Song (2025)	Pose estimation in complex dance scenes	Enhanced HRNet backbone with improved feature extraction and robustness modules for cluttered scenes	Shows improved keypoint accuracy and better robustness under occlusion/clutter.
Büyükgökoğlan and Uğuz (2025)	Deep-learning–based scoring for Turkish folk dance	Webcam capture; pose extraction via MediaPipe/YOLO; comparative models: DTW, TLCC, LSTM, Siamese networks	LSTM produced higher scores (68.43, MSE 56.11) vs DTW (60.64, MSE 139.32); system sensitive to camera angle.

2.2 Lacking in the Approaches

These studies show the potential of pose estimation and deep learning for advancing both modern dance movement design and traditional folk dance preservation. However, there is little to no research in the Philippines that applies pose estimation to folk dances—particularly Tinikling—representing a significant gap and opportunity for future exploration.

2.3 Summary

Research on human pose estimation (HPE) spans multiple applications including AR/VR, healthcare, and dance. Optimization- and deep learning–based models (e.g., SMPL, SMPL-X, HMR, VIBE, SPIN, PARE, EXPOSE, PHALP) have been studied for realistic 3D body reconstruction (Venkatrayappa et al., 2024). Dance classification has been explored using skeleton data and machine learning classifiers like k-NN and Random Forest (Protopapadakis et al., 2018). Transformer-based models such as DanceFormer achieve high accuracy and real-time performance in dance pose estimation (Zhao et al., 2025), while PAF- and LSTM-based algorithms improve movement recognition (Lei et al., 2023). Kinect studies reveal both potential and limits in low-cost motion capture (Tölgyessy et al., 2021; Rallis et al., 2019), while feedback and sequence-alignment approaches (Lin et al., 2015; Yu Xiong, 2019) highlight the importance of interactivity and temporal robustness.

Recent work integrates advanced neural networks for pose estimation, such as adaptive hypergraphs (Xu et al., 2022), deep feature fusion for dance poses (Ju, 2025), MoCap for authentic folk dance templates (Rizhan et al., 2025), and deep learning systems for evaluating Turkish folk dance (Büyükgökoğlan Uğuz, 2025).

Chapter 3

THEORETICAL CONSIDERATIONS

Chapter 4

DESIGN CONSIDERATIONS

Chapter 5

METHODOLOGY

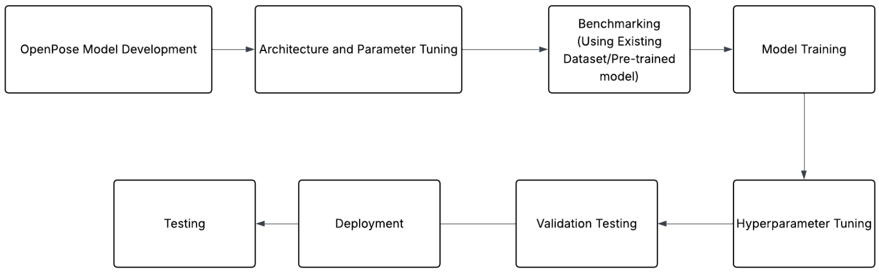

Fig. 5.1 Methodology Flowchart

5.1 Methodology

5.2 Design Considerations

5.2.1 Sensor choice, representation, and robustness

A study by Tölgyessy, Dekan, and Chovanec (2021) demonstrated that Kinect-family depth sensors produce explicit 3D skeletons and give higher joint fidelity in controlled settings, but the accuracy falls with occlusion, off-axis views, and increased distance. Zhang et al. (2020) described MediaPipe, which yields compact 2D/3D landmark coordinates from ordinary RGB cameras and runs in real time on mobile devices. Therefore, designers often choose landmarks for rapid, lightweight prototypes and mobile deployment, and reserve depth or IR systems for installation-grade fidelity when hardware is available. To reduce real-world failure modes, practitioners apply photometric and background augmentation and synthetic occlusions during training, and they add a short calibration step so system metrics align with an individual user's range of motion.

5.2.2 Temporal alignment and scoring

Dance is a temporal activity and should be compared as a sequence rather than as isolated frames. Yu and Xiong (2019) demonstrate that Dynamic Time Warping (DTW) can align noisy, tempo-varying Kinect skeleton sequences and convert DTW distances into meaningful performance scores. Rallis et al. (2019) apply DTW to choreographic trajectories and show it can match patterns across high-precision (VICON) and low-cost (Kinect) capture systems. Thus, a practical scoring pipeline first aligns sequences with DTW (or a constrained variant) and then evaluates local spatial metrics such as joint-angle differences or normalized trajectory distances to produce interpretable, per-segment correctness scores.

5.2.3 Real-time feedback, segmentation, and pedagogy

Lin (2015) finds that immediate, clear feedback in dance exergames improves engagement and supports learning. Zhang et al. (2020) show that on-device landmark extraction can run at real-time rates suitable for low-latency feedback. Combining these results suggests a two-tier runtime design: use a fast, coarse matcher (enabled by on-device landmarks) for instant cues, and run a slower, higher-precision alignment and scoring pass for final grading. Breaking choreography into short labeled segments also simplifies alignment and reduces error accumulation; Rallis et al. (2019) illustrate that segment- or trajectory-level matching better supports choreographic retrieval and per-segment feedback.

5.2.4 Accessibility, personalization, and evaluation

Yu and Xiong (2019) convert DTW distances into calibrated percentage scores, which supports per-user calibration and comparison against an individualized baseline. Tölgyessy et al. (2021) recommend measuring sensor-level metrics such as joint error and dropout rates when choosing a capture modality. Therefore, system designs should include adjustable sensitivity, alternate gesture mappings, and user profiles, and evaluation should combine sensor metrics (joint error, dropout, latency) with human-centered measures (perceived accuracy, engagement, and learning gain) to justify architecture and scoring choices.

5.3 Theoretical Considerations

5.3.1 Human Pose Estimation

Human pose estimation is the process of predicting the pose of human body parts. The data are typically stemming from RBD images or videos. Given that certain motions are motivated by human actions, detecting poses is a critical aspect of human action recognition (Song et al., 2021). It has a wide range of applications such as human-computer interaction, motion analysis, augmented reality, and virtual reality. The resulting output of human pose estimation is a skeleton-like representation of the human body consisting of nodes and limbs (Zheng et al, 2020)). There are 2 main types of human pose estimation, namely 2D and 3D. 2D pose estimation consists of predicting the posture of each of the body's key points in a 2D plane, considering the X and Y axis. As for 3D pose estimation, it considers the Z axis, situating each point in a 3D space. It goes without saying that the 3D estimation would be much more difficult in comparison to 2D estimation in process or complexity due to underlying issues which may manifest such as noisy backgrounds, clothing, lighting, undetected joints, or occlusion (Ben Gamra Akhloufi, 2021).

5.3.2 Human Action Recognition

Human action recognition, otherwise known as HAR. is the process of detecting human actions in order to classify them through single sensor data, RBD image or video data, or three-dimensional depth and inertial data (Sakar et al., 2022). In the field of computer vision, one of the most challenging aspects of it is the automatic and precise identification of human activity. Over the years, there has been a significant increase in feature learning-based representations for human action recognition as a result of the widespread utilization of deep learning-based features. There are various applications of Human action recognition. For instance, automated surveillance systems make use of AI and machine learning algorithms in order to identify human actions for the sake of safety and security. Such a task, however, is made difficult due to various factors such as changing online environments, occlusion, different viewpoints, execution pace and biometric change. Not only this, but the human body also varies from person to person in factors such as size, appearances, and shapes. However, advancements in Convolutional Neural Networks, otherwise known as CNNs, resulted in significant progress for human action recognition through improvements on classification, segmentation and object detection. This largely applies more on image-related tasks rather than videos as neural network models struggle to capture temporal information in videos due to a lack of substantial datasets (Morshed et al., 2022).

5.4 Summary

Provide the gist of this chapter such that it reflects the contents and the message.

REFERENCES

Büyükgökoğlan, E. and Uğuz, S. (2025). Development of a performance evaluation system in turkish folk dance using deep learning-based pose estimation. *Tehnicki vjesnek - Technical Gazette*, 32:1817–1824.

Lei, P., LI, N., and Liu, H. (2023). *Dance Movement Recognition Based on Gesture*, pages 448–452.

Lin, J.-H. (2015). "just dance": The effects of exergame feedback and controller use on physical activity and psychological outcomes. *Games for Health Journal*, 4(3):183–189. PMID: 26182062.

Oudah, M., Al-Naji, A. A., and Chahl, J. (2020). Hand gesture recognition based on computer vision: A review of techniques. *Journal of Imaging*, 6:73.

Protopapadakis, E., Voulodimos, A., Doulamis, A., Camarinopoulos, S., Doulamis, N., and Miaoulis, G. (2018). Dance pose identification from motion capture data: A comparison of classifiers. *Technologies*, 6.

Rahman, M. M., Uzzaman, A., Khatun, F., Aktaruzzaman, M., and Siddique, N. (2025). A comparative study of advanced technologies and methods in hand gesture analysis and recognition systems. *Expert Systems with Applications*, 266:125929.

Rallis, I., Protopapadakis, E., Voulodimos, A., Doulamis, N., Doulamis, A., and Bardis, G. (2019). Choreographic pattern analysis from heterogeneous motion capture systems using dynamic time warping. *Technologies*, 7(3).

Sun, J. and Song, L. (2025). Dance movement pose estimation in complex scenes based on improved high-resolution networks. *Taiwan Ubiquitous Information*, 10(1):—.

Tölgyessy, M., Dekan, M., and Chovanec, (2021). Skeleton tracking accuracy and precision evaluation of kinect v1, kinect v2, and the azure kinect. *Applied Sciences*, 11(12).

Ultraleap (2025). Ultraleap hand tracking overview.

Venkatrayappa, D., Tremeau, A., Muselet, D., and Colantoni, P. (2024). Survey of 3d human body pose and shape estimation methods for contemporary dance applications.

Verma, A. R., Singh, G., Meghwal, K., Ramji, B., and Dadheech, P. K. (2024). Enhancing sign language detection through mediapipe and convolutional neural networks (cnn).

Xu, X., Zou, Q., and Lin, X. (2022). Adaptive Hypergraph Neural Network for Multi-Person Pose Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2955–2963.

Yu, X. and Xiong, S. (2019). A dynamic time warping based algorithm to evaluate kinect-enabled home-based physical rehabilitation exercises for older people. *Sensors*, 19(13).

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C., and Grundmann, M. (2020). Mediapipe hands: On-device real-time hand tracking. *CoRR*, abs/2006.10214.

Zhao, H., Du, B., Jia, Y., and Zhao, H. (2025). Danceformer: Hybrid transformer model for real-time dance pose estimation and feedback. *Alexandria Engineering Journal*, 121:66–76.

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., and Shah, M. (2023). Deep learning-based human pose estimation: A survey.

Chapter F
MEMBER SKILLSET IDENTIFICATION

TABLE F.1 TEAM MEMBERS' PROGRAMMING SKILLS

Member	**Model Dev.**	**UI Design**	**Source Control (GitHub)**	**Problem Solving & Opt.**	**Python**
Hans	Intermediate	Novice	Expert	Intermediate	Intermediate
Gerald	Intermediate	Basic	Novice	Intermediate	Intermediate
Nathan	Intermediate	Novice	Novice	Intermediate	Intermediate

Chapter G WORK BREAKDOWN STRUCTURECAPSTONE PROJECT ON OPERATIONAL TECHNOLOGIES

De La Salle University

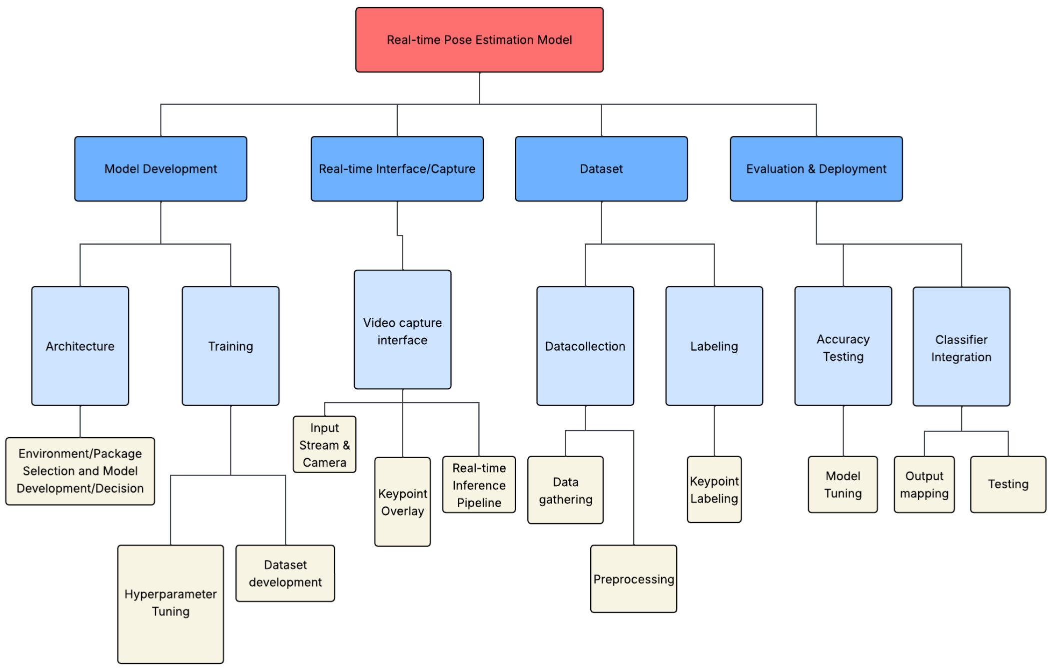

Fig. G.1 Work Breakdown Structure for Hans Capstone Project on Operational Technologies

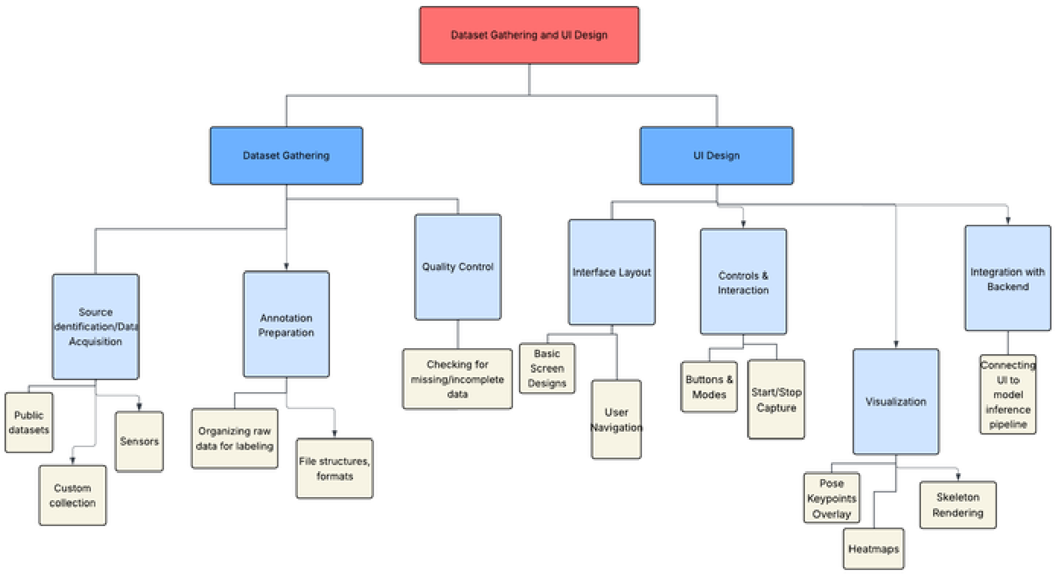

Fig. G.2 Work Breakdown Structure for Nathan Capstone Project on Operational Technologies

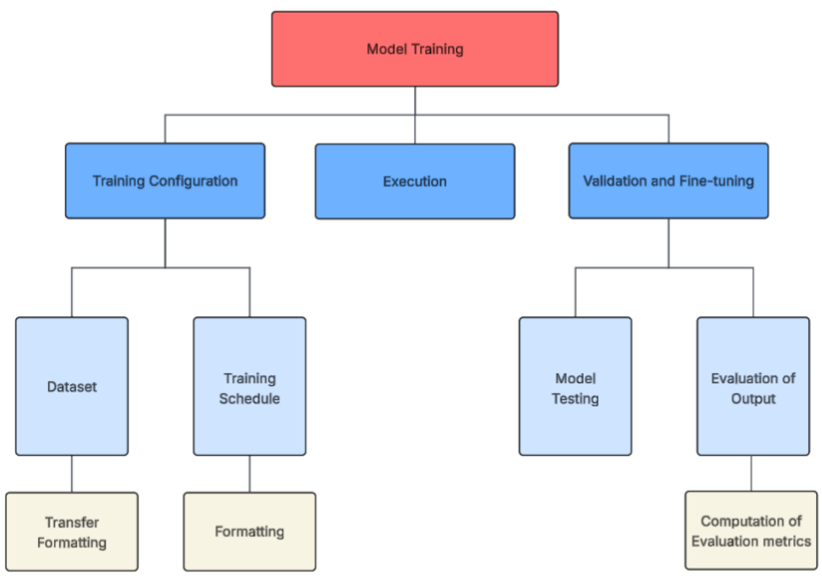

Fig. G.3 Work Breakdown Structure for Gerald Capstone Project on Operational Technologies