SayawMo:
A Recognition Model for Western-Influenced Philippine Folk Dance

Herbieto, Kermichil A.*[1], and Roxas, Robert R.[2]

[1,2]University of the Philippines Cebu, Cebu City, Cebu, Philippines
[1]kaherbieto@up.edu.ph
[2]robert.roxas@up.edu.ph

Abstract. Folk Dance showcases both the performative and cultural dimensions of dance. Although its instruction has been integrated into the Philippine education system, the declining interest and exposure, compounded by the challenges of instruction and lack of surrounding research, prompts the need for alternate solutions for dance representation. The integration of technology in the realm of dance presents a transformative opportunity to explore and support various perspectives of dance, in education, or historical preservation. These can be accomplished through developing a dance model. Therefore, a supervised dance recognition model for the upper and lower body dance terms in Western-influenced Philippine Folk Dance was built using the Long Short-Term Memory (LSTM) Neural Network alongside the MediaPipe pose model. While the upper body model achieved an F1 score of 0.9188, effectively representing the upper body dance literature, the lower body model struggled with an F1 score of 0.2284, failing to adequately represent the lower body dance literature.

Keywords: Long-short Term Memory, MediaPipe, Western-influenced Philippine Folk Dance, Pose Model

1 INTRODUCTION

The application of dance to technology can provide the resources to support the art of dancing in the exploration of its different fields and perspectives. Technological developments with dance applications allow the representation of dance movements as data [1]. Dance recognition presents a novel method for teaching the performance of dance while serving as a tool for researching new dance methodologies [2]. This is more prominent in contemporary dance research that focuses on the combination of dance generation to music, setting movement to the individual frames of sound [3]. As such, the intersection between dance and technology can contribute to a further understanding of the art form, especially within the fields of e-learning and historic preservation [1].

Folk Dance is the dance that captures the essence of a community's culture and is performed either on festive occasions or in performance shows. In the Philippines, the knowledge of folk dance has been integrated into the education system, forming a link between education, culture, and dance [4]. The rising modernity and the accessibility of expression that contemporary dance, however, can offer led to a decrease in both interest and exposure to folk dances. The issue remains especially difficult due to the complexity of the dances, which requires additional material to aid instruction [5]. This highlights the need for a way to not only make the demonstration of folk dances accessible but also in the development of a tool that can aid researchers, educators, and

J. Caro et al. (eds.), Proceedings of the Workshop on Computation: Theory and Practice (WCTP 2024) , Atlantis Highlights in Computer Sciences 23,
https:// doi.org/10.2991/978-94-6463-684-0_7

performers of folk dance. With this, there is a need to revitalize the program of Philippine folk dances from academic to cultural development [6].

This research addresses the need to preserve Philippine folk dances by leveraging machine learning for dance recognition. The study aims to develop a model capable of identifying and understanding the fundamental movements characteristic of Western-influenced Philippine folk dances. By accurately capturing these dance features, the model will serve as a valuable tool for educators and performers, aiding in the teaching and preservation of these cultural practices. Additionally, this project establishes a foundational framework for integrating technology with the traditional art of folk dance.

2 REVIEW OF RELATED LITERATURE

This research involved the examination of the machine learning model applied to the problem, and along with its application to pose recognition. This contributed to creating a structured framework on the approach towards this research.

2.1 Neural Networks for Temporal Data

There are two common Neural Network models used in projects and research concerning video data: Convolutional Neural Network (CNN) and Recurring Neural Network (RNN). CNN is developed to recognize visual patterns from image pixels [7] which is more appropriate for tasks involving spatial data, like images. It is designed to operate on inputs and generate outputs of fixed dimensions, provide predictions at the appropriate level, and include a confidence measure for their predictions. RNN is commonly employed for handling sequential data, treating each frame of the video as an input. The work highlighted in [8] highlighted that CNN performed better when applied to image data than RNN. For tasks that involve video data, RNN performed better than CNN [7]. One interesting observation in their paper was that the combination of both CNN and RNN results in a model that yields a high accuracy. Long Short-Term Memory (LSTM) is a type of RNN that exists to solve the problem of vanishing gradients or long-term dependence that exists in traditional RNN models, especially for datasets that contain relatively long data [3][9]. This was corroborated by the results reported in [10], where the CNN-LSTM model performed the best among models that utilized the CNN and LSTM.

2.2 Human Pose Libraries

The work in [11] compared the capabilities of the MediaPipe, MoveNet, OpenPose, and PoseNet pose estimation libraries. In this study, the Microsoft Common Object in Context (COCO) dataset and the Penn Action datasets were used to capture the capabilities of these libraries in the application of both image and video data. In their findings, the OpenPose library has the lowest performance when applied to video data, yet it works second best when applied to image data. Therefore, they recommended the use of the MoveNet pose library for both image and video recognition applications. This was corroborated by the investigation the work in [12] regarding pose estimation of the MoveNet, MediaPipe, and PoseNet libraries. They concluded that MoveNet performed

the best in the COCO dataset and both MoveNet and MediaPipe worked best in a live environment implementation.

The MoveNet and MediaPipe pose libraries were both considered for the study. MoveNet employed a top-bottom approach, detecting a human figure first before deriving its nodes, while MediaPipe used a bottom-up approach, detecting all nodes first and then inferring their relationships [11][13]. MediaPipe covers the key points detected by the MoveNet pose library while capturing additional key points for hands and feet as shown in Figure 1. This made it particularly suitable for tasks that required contextual information about hand and foot positions.

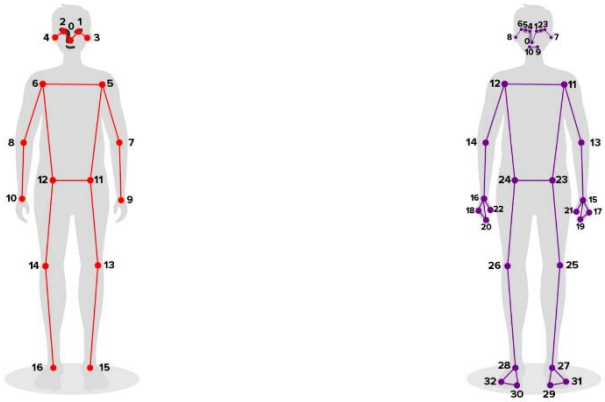

Fig. 1. Key points detected in MoveNet (left) and MediaPipe (right) pose libraries

2.3 Pose Recognition Using Pose Libraries

Bowling Deep-Learning (BDL) utilized the MoveNet model for detection, and the dataset that the model generated was sent to the BDL for classification. This provided the approach to reconcile the machine-learning model and pose library in the application of pose recognition. The process involved the combination of skeletonization through the use of the MediaPipe pose model and the processing of the information through the MoveNet model, as MoveNet uses a CNN trained in poses. [14] Although the paper discusses the application of pose recognition concerning image data, it still achieves the proposed solution of bridging pose library and a machine learning model.

The work in [15] developed a dance recognition model that would recognize basic hand poses in classical Chinese dances using a cloud AI and the OpenPose library. This provided a framework that the paper was able to capture the basic pipeline for a recognition process and how labeling is done towards the pose library in collaboration with its classification model. Despite this, the paper did not include its results nor the dataset it is using. Any further analysis of the effectiveness of the model was difficult to conclude.

Further development in the models relies on the handling of the data before the application of the pose models. The work in [16] used the MoveNet model in pose recognition, while video data was extracted through the Exponentially Reduced Ordinal Surface (EROS), where performance is at a competitive state with its contemporaries.

3 METHODOLOGY

Video data were collected from YouTube videos titled "Dance Terms in Philippine Folk Dance" or similar structured titles. These videos were manually segmented and classified according to their movements. The raw segmented video data was then applied to the MediaPipe pose model, where the data was transformed into a standardized array stored in JSON file with a hierarchy of sample, frame number, and flattened dimension value. The data was further processed into a uniform number of frames to accommodate the performance of the LSTM network. The model was then built, tested for its predictions, and evaluated.

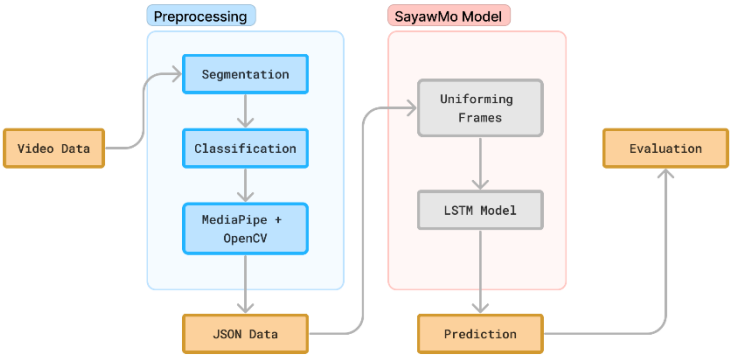

Fig. 2. The design framework of the study.

Initial data gathering involved searching the YouTube website for every applicable video for Fundamental Figures in Philippine Folk Dance. Once enough video data were gathered, they were downloaded with the yt-dlp downloader. The total duration of the gathered videos used in this study amounts to one (1) hour, two (2) minutes, and 29 seconds.

The videos were segmented into parts, where only one human subject was present in the video and the subject would perform only one movement. The videos were then classified according to the label tagged by the video owner, while ambiguous data were corroborated with the UP Iskultura Dance Ensemble for proper identification with the proper literature of the movement. Videos were segregated according to upper or lower model classification.

The videos underwent processing to the MediaPipe pose model through the OpenCV library. They were cropped to isolate the human subject. For each frame, the model extracted x, y, and z values from the recognized joints, which were then flattened and added to the JSON file. A separate JSON file containing the labels for each sample was also created alongside the dataset.

The generated dataset was processed into the SayawMo model, which was created through the Tensorflow-Keras library in Python. The architecture of the model is shown in Table 3.1, where it consisted of three LSTM and two dropout layers alternating with one another. Two dense layers were found at the end to narrow the model's decision to one category. For every epoch generation, the dataset was split into a ratio of 70%

testing data and 30% validation data. The accuracy, validation accuracy, loss, and validation loss, along with its confusion matrix, were tracked throughout each epoch generation. Numerical analysis was performed with the Sklearn library and visual analysis was performed with the aid of the Bokeh library.

4 RESULTS AND DISCUSSION

This section detailed the findings of the experiment done in the Methodology chapter in terms of the upper and lower model datasets.

4.1 Upper Body Dataset

From the videos that were gathered, the upper model dataset consisted of nine (9) categories and had an overall duration of 16.73 minutes. The 'Lateral, 'Hayon-Hayon' and 'Sarok were the common poses that were manually classified from the videos, with a density of 16.52%, 15.41%, and 13.29%, respectively. 'Arms in Reverse "T"' was the category that had the lowest duration in the dataset having a density of 2.25%.

The model was evaluated for its performance and its values were recorded as shown in Table 1. It can be observed from Figure 3 that the accuracy would dip on certain epochs. The general trend of the model would indicate that the model had captured the general features of the dataset as its accuracy and validation accuracy contained relatively small distances between each other in each epoch generation. Despite this, the accuracy and validation accuracy would slowly diverge as the epoch increases. This would infer that the model would slowly overfit with the current dataset and would continue to perform so with increased epochs.

Table 1. Upper model evaluation

Category	Value
Accuracy	0.85465
Validation Accuracy	0.70667
Loss	3.48329
Validation Loss	4.10537
Precision	0.89505
Recall	0.84367
F1	0.84821

The model, as shown in Figure 4, was able to achieve an F1 score of 0.8482, indicating that it was able to correctly predict most of the dataset. Although these values would be promising, an observation between the model evaluation and the learning performance would imply that the dataset was not coherent enough as training data as indicated by the spikes of accuracy. Still, the model would be expected to grow properly since the accuracy and F1 score were similar.

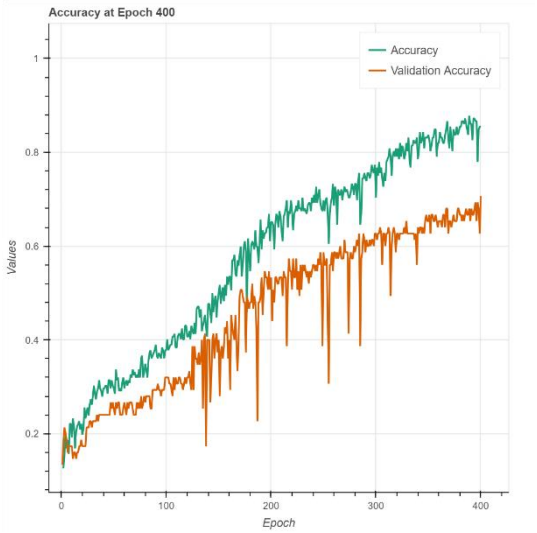

Fig. 3. Accuracy (green) and validation accuracy (red) of the upper body dataset at every epoch

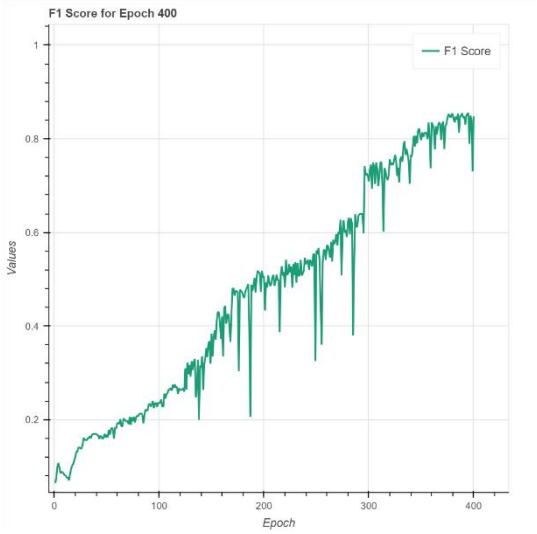

Fig. 4. F1 score of the upper body dataset at every epoch

Poses with similar movement forms, such as 'Masiwak' and 'Kumintang', and 'Sarok', 'Salok', and 'Bilao' were predicted correctly, indicating that the model could learn complex data from the dataset. An exception to this is the 'Patay' pose which posed confusion with the performance of the model. It was observed that none of the 'Patay' data were included in the evaluation, even throughout many different trials. This indicated that there was an issue with the 'Patay' video data that the preprocessing step ignored the data. Video quality could influence the ability of the MediaPipe pose model to properly detect the nodes.

4.2 Lower Body Dataset

The lower model dataset consisted of 16 categories and had an overall duration of 19.62 minutes. ‘Hop’, ‘Point’, and ‘Brush’ were the most represented categories, where their dataset density amounted to 10.72%, 9.90%, and 9.82%, respectively. ‘Step’ had the lowest duration in the dataset at 1.86%, followed by 'Jump' and 'Draw' with 2.64% and 3.03%, respectively.

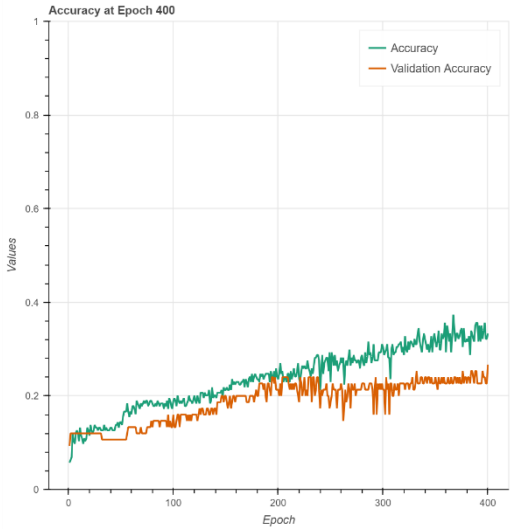

Fig. 5. Accuracy (green) and validation accuracy (red) of the lower body dataset at every epoch

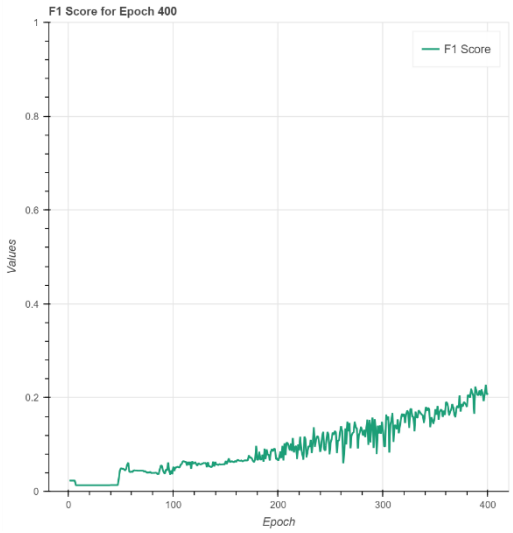

Fig. 6. F1 score of the lower body dataset at every epoch

Through the results of the model, as indicated by Table 2, it achieved an F1 score of 0.2284. This value indicated that the model was not able to successfully capture the features of the lower dataset. The validation accuracy of the model began to diminish

around the 300th epoch, as shown in Figure 5, which contributed to its low F1 score compared to its accuracy, as shown in Figure 6. These factors imply that the lower model was not able to represent the lower dataset of the study.

Table 2. Lower model evaluation

Category	Value
Accuracy	0.33333
Validation Accuracy	0.26667
Loss	4.65302
Validation Loss	4.90792
Precision	0.27205
Recall	0.28028
F1	0.22835

The model could correctly predict the 'Draw' and 'Waltz' movements. There were multiple movements that the model did not predict correctly, which correlated to the low model accuracy and F1 score. These movements were 'Step', 'Stamp', 'Panandyak', 'Leap', 'Brush', 'Bleking', and 'Point'. This could indicate that even with three LSTM layers, the lower dataset was too complex to capture and would explain the poor performance. Poor video quality, as described in the upper model, could also explain the poor performance.

5 CONCLUSIONS AND RECOMMENDATIONS

The study was able to partially accomplish the objective of building a model that would recognize dance figures in Western-influenced Philippine Folk Dance. It was seen that the upper model performed well, with it being able to capture the features of the upper model dataset. The lower model was not able to perform effectively, having an incredibly low accuracy and F1 score. With the experiment results, the SayawMo model is not ready to be integrated into dance applications without resolving the lower model. Nevertheless, the study was able to establish a framework to develop a dance model from dance terms found in Philippine Folk Dance.

An improved dataset will greatly benefit the entire study, as although video data gathered from video hosting sites are convenient, there are few credible resources for dance terms in Philippine Folk Dance. Even if there are videos for the dance terms, the dataset will be skewed towards particular terms than others as indicated by the lower model dataset. The improved dataset should be gathered with the collaboration of a dance ensemble for Folk Dance to not only establish credibility with the data but also control the density of the dataset.

Hyperparameter optimization was unexplored during the experimentation, possibly resulting in a less efficient model that could capture the dance features. Thus, further exploration of the LSTM architecture and its settings would benefit the prediction of the lower dataset due to the complexity of capturing legwork in dance.

Additional features, such as velocity and position relative to the inner and outer foot or arm, could be added to the current MediaPipe pose model. This may contribute positively to the performance of the lower model, accommodating its complexity.

Lastly, an implementation of an application for dance recognition should be included in future iterations of the study to assess the model's performance in real scenarios, also in collaboration with a Folk Dance ensemble. This method would allow further evaluation of the SayawMo model in terms of correctly capturing the features of folk dance, as well as its application for education and historic preservation.

6 IMPLICATIONS

This study serves as a preliminary model that will be used in the development of future studies involving Philippine folk dances, especially in the context of pose recognition and potentially pose generation. This would also serve as a framework for pose recognition models that utilize pose libraries and machine learning, as well as an investigation concerning the quality and availability of training materials. This study also serves to expand the application of the MediaPipe pose model as the input for the LSTM model, particularly toward Western-influenced Philippine Folk Dances.

7 ACKNOWLEDGEMENT

The author would like to acknowledge Robert Roxas, the research adviser, for guiding and refining the methodology of the study. The author would also like to acknowledge Jayvee Opina and Javy Manayon, dance coaches for Philippine Folk Dance, for corroborating the dance literature used in building the model.

References

1. Joshi, M. & Chakrabarty, S.: An extensive review of computational dance automation techniques and applications. In: Proceedings of the Royal Society: A Mathematical, Physical and Engineering Sciences (2021).
2. Bisig, D.: Generative dance – A taxonomy and survey. In: Proceedings of the 8th international conference on movement and computing. Association for Computing Machinery, New York, NY, USA (2022).
3. Zaman, L., Sumpeno, S., Hariadi, M., Kristian, Y., Setyati, E., & Kondo, K.: Modeing basic movements of Indonesian traditional dance using generative long short-term memory network. In: IAENG International Journal of Computer Science, 47(2), pp. 262-270 (2020).
4. Santos, M. F.: Philippine folk dances: A story of a nation. In: Journal of English Studies and Comparative Literature, 18(1), pp. 4-6 (2019).
5. Lobo, J.: Protecting Philippine dance traditions via education of tomorrow's pedagogues: The role of individual interest and school engagement. In: Journal of Ethnic and Cultural Studies, 10(1), pp. 98-124 (2023).
6. Domingo, J. R.: Articulating Philippine folk dance documentation practices. In: De La Salle University Research Congress, BUILDING IMPACT ON FIRM FOUNDATIONS: FROM BASICS TO APPLICATIONS, pp. 4 (2018)
7. Singhal, N., Singhal, N. & Srishti.: Comparing CNN and RNN for prediction of judgement in video interview based on facial gestures. In: 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 175-180 (2018)

8. Qing, X.: A comparison study of convolutional neural network and recurrent neural network on image classification. In: Proceedings of the 2022 10th International Conference on Information Technology: IoT and Smart City. Pp. 112-117 (2022)
9. Xiao, H., Sotelo, M. A., Ma, Y., Cao, B., Zhou, Y., Xu, Y., & Li, Z.: An improved LSTM model for behavior recognition of intelligent vehicles. In: IEEE Access, pp. 124928-124938 (2020)
10. Wang, S., Li, J., Cao, T., Wang, H., Tu, P., & Li, Y.: Dance emotion recognition based on laban motion analysis using convolutional neural network and long short-term memory. In: IEEE Access 8, pp. 101514-101527 (2020)
11. Chung, J. L., Ong, L. Y., & Leow, M. C.: Comparative analysis of skeleton-based human pose estimation. In: Future Internet, 14(12) pp. 380 (2022)
12. Bolaños, C., Fernandex-Bermejo, J., Dorado, J., Agustin, H., Villanueva, F. J., & Santofimia, M. J.: A comparative analysis of pose estimation models as enablers for a smart-mirror physical rehabilitation system. In: Procedia Computer Science, 207, pp. 2536-2545 (2022)
13. BeomJun, J., & SeongKi, K.: Comparative analysis of OpenPose, PoseNet, and MoveNet models for pose estimation in mobile devices. In: Traitement Du Signal, 39(1), 119-124. (2022)
14. Janbi, N. F., & Almuaythir, N.: BowlingDL: A deep learning-based bowling players pose estimation and classification. In: 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), pp. 1-6 (2023)
15. Zhou, Q., Wang, J., Wu, P., & Qi, Y.: Application development of dance pose recognition based on embedded artificial intelligence equipment. In: Journal of Physics: Conference Series, 1757(1) (2021)
16. Goyal, G., Pierto, F. D., Carissimi, N., & Glover, A.: MoveEnet: Online high-frequency human pose estimation with an event camera. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4024-4033 (2023)