Indian Institute of Information Technology Vadodara
(Gandhinagar Campus)

Design Project Report-2023
on
Yoga Pose Detection Using LSTM

Mentor: Dr. Jignesh Bhatt
Team Members:
Ayush Pathak (202151193)
Subham Rathi (202151163)
Sumit Kumar (202151165)
Shivam Pawar (202151148)

***Abstract*—Yoga Pose Detection Using LSTM is a machine learning project that focuses on the application of Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) for the detection of yoga poses in videos. The research aims to develop a system capable of analyzing video sequences to identify and classify various yoga poses accurately. The integration of LSTM and CNN is explored for its potential in capturing temporal dependencies and spatial features in the video data. The report presents a detailed analysis of the methodology, results, and discussions, along with conclusions and suggestions for future work.**

I. Introduction

Yoga, as a form of physical and mental exercise, has gained widespread popularity. Automating the detection of yoga poses holds great potential for monitoring and improving practice. This project employs machine learning techniques, specifically Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), to automate the detection of yoga poses in videos.

The increasing popularity of yoga has led to a growing interest in leveraging technology for enhancing the practice experience. Our project aims to contribute to this intersection of yoga and technology by developing an intelligent system capable of recognizing and analyzing yoga poses in video recordings.

II. Literature Review

Previous studies in the field of pose detection and human activity recognition have explored various methods. The integration of LSTM for capturing temporal dependencies and CNN for spatial features has shown promising results. Researchers have applied similar techniques to recognize complex movements and activities.

In our literature review, we delve into the existing methodologies and technologies employed in related projects. We explore the strengths and limitations of different approaches, providing a comprehensive understanding of the current landscape of yoga pose detection using machine learning.

III. Methodology

The methodology involves several key steps, including data preprocessing, model training, and evaluation. Video frames are extracted, normalized, and augmented to create a robust dataset. The LSTM-CNN model is then trained on this dataset to recognize and classify yoga poses. The training process includes fine-tuning the model's weights to enhance accuracy.

The dataset used in our project consists of diverse yoga poses performed by individuals of varying skill levels. Each video undergoes preprocessing to extract relevant frames, which are then used to train the machine learning model. We utilize transfer learning techniques to leverage pre-trained models and optimize the performance of our system.

A. Data Preprocessing

A research paper featured in Neural Computing and Applications detailed the creation of a structured dataset characterized by a consistent resolution of 1280x720 pixels and a standard frame rate of 30 frames per second. This dataset focused on documenting six distinct yoga poses, each comprising 25-second video clips recorded by six individuals for each pose.

During the preprocessing phase, the team amalgamated the videos corresponding to each pose while eliminating any extraneous noise. These initial videos were trimmed to a duration of 2.5 minutes for each pose. Subsequently, these videos were further segmented into sequences consisting of 125 frames, aligning with the objective of capturing a 5-second clip at a frame rate of 25 frames per second.

Consequently, the finalized dataset comprised six primary directories, each dedicated to a specific yoga pose. Within these directories were 24 subdirectories, each containing sequences of 125 frames, representing distinct recordings of the yoga poses captured by various individuals.

B. Model Training

Our model leverages the power of LSTM and CNN architectures. LSTM networks capture temporal dependencies in sequential data, making them well-suited for analyzing video frames over time. CNNs excel at extracting spatial features, aiding in the identification of key pose characteristics.

Transfer learning is employed to take advantage of pre-trained models on large image datasets. This approach allows our model to benefit from the knowledge gained during the training of these models on extensive image datasets.

ANN Model

The ANN model is a deep learning model tailored for classification tasks. Constructed using the Keras Sequential API, it features four dense layers for applying linear transformations to the input. BatchNormalization layers normalize activations, and Dropout layers mitigate overfitting by randomly dropping input units. Boasting 45,126 parameters, the ANN model excels in accuracy-demanding classification tasks like image classification and sentiment analysis. It has proven its mettle with state-of-the-art performance on various benchmark datasets

25fps Model

The 25fps model is a deep learning model designed for video processing tasks. Utilizing the Keras Sequential API, the model comprises three LSTM layers for processing sequential data, such as video frames. Dropout layers are strategically placed to prevent overfitting, while BatchNormalization layers normalize activations at each batch. With a total of 334,598 parameters, this model is particularly beneficial for video processing tasks demanding high temporal resolution, including action recognition and video captioning. It has demonstrated state-of-the-art performance on multiple benchmark datasets

10fps Model

The 10fps model, another deep learning solution for video processing tasks, is defined using the Keras Sequential API and comprises three LSTM layers. Similar to the 25fps model, Dropout and BatchNormalization layers are employed to prevent overfitting and normalize activations, respectively. With 334,598 parameters, the 10fps model is ideal for video processing tasks requiring high temporal resolution, such as action recognition and video captioning. Its state-of-the-art performance has been validated across multiple benchmark datasets

C. Evaluation

The model's performance is evaluated using standard metrics such as accuracy, precision, recall, and F1 score. The evaluation phase helps validate the effectiveness of our approach and identify areas for improvement.

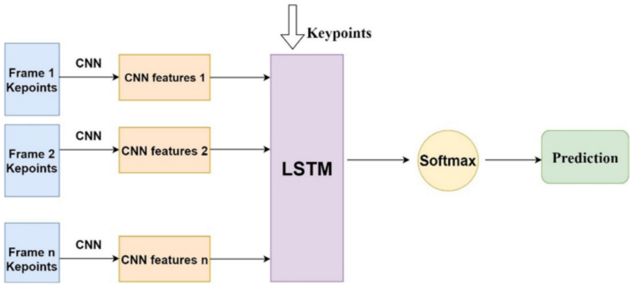

Fig. 1. System architecture: OpenPose followed by CNN and LSTM model

IV. Results and Discussions

The results section presents a thorough evaluation of the investigation, showcasing the contributions of the study. The discussion covers inferences drawn from the results, conclusions, and potential avenues for further research. Challenges encountered during implementation are addressed, providing valuable insights.

A. Model Accuracy Graph

A key visual representation of our results is the model accuracy graph, depicting the performance of our LSTM-CNN model over training epochs. This graph illustrates the convergence of the model and highlights any potential overfitting or underfitting.

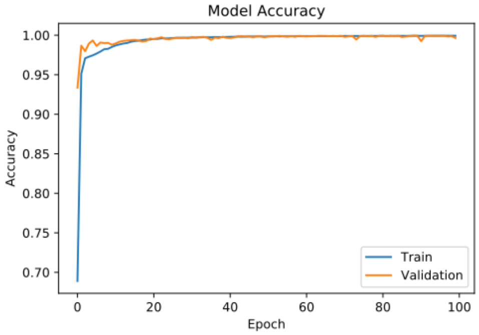

Fig. 2. Model Accuracy Over Training Epochs.

B. Confusion Matrix (Frame-Based)

To further analyze the model's performance for frame-based detection, a confusion matrix is presented. This matrix provides insights into the model's ability to correctly classify different yoga poses and identifies common misclassifications.

C. Confusion Matrix (Sequence-Based)

For sequence-based detection, another confusion matrix is presented. This matrix focuses on the model's ability to recognize patterns in sequences of frames, providing a more comprehensive understanding of its performance.

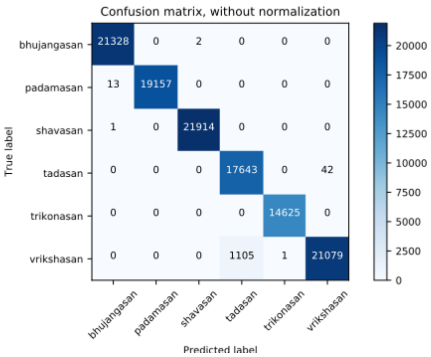

Fig. 3. Confusion Matrix (Frame-Based) Illustrating Model Performance.

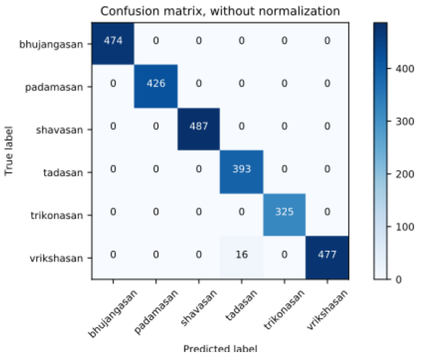

Fig. 4. Confusion Matrix (Sequence-Based) Illustrating Model Performance.

D. *Qualitative Results*

In addition to quantitative metrics, qualitative results are showcased through visual examples. Snapshots from the model's predictions on sample yoga pose videos highlight its ability to accurately identify and classify poses.

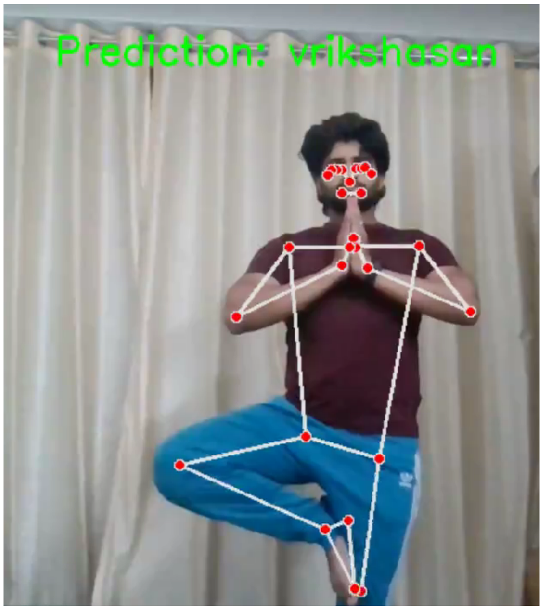

Fig. 5. Qualitative Results of Yoga Pose Detection.

V. Conclusions and Future Work

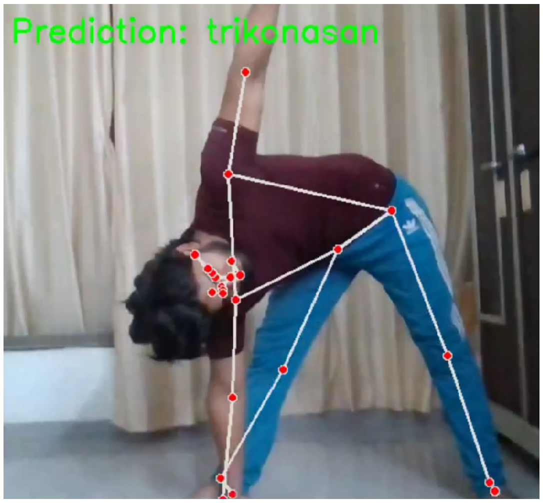

Fig. 6. Qualitative Results of Yoga Pose Detection.

The concluding section summarizes the key findings and insights derived from the study. The project successfully demonstrates the feasibility of using LSTM and CNN for yoga pose detection. Future work could involve expanding the dataset, refining the model architecture, and exploring real-time applications.

A. *Challenges and Lessons Learned*

Reflecting on the challenges encountered during the project, we identify key lessons learned. Addressing issues related to dataset diversity, model complexity, and real-world applicability are crucial for the advancement of yoga pose detection systems.

VI. Acknowledgment

The authors express gratitude to **Dr. Jignesh Bhatt** for his guidance and support throughout the duration of this project. Dr. Bhatt's expertise in machine learning and computer vision significantly contributed to the success of our research.

VII. References

1) Islam, Muhammad Usama, et al. ”Yoga posture recognition by detecting human joint points in real time using Microsoft Kinect.” 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). IEEE, 2017.
2) Chen, Hua-Tsung, et al. ”Yoga posture recognition for self-training.” International Conference on Multimedia Modeling. Springer, Cham, 2014.
3) Jin, Xin, et al. ”Virtual personal trainer via the Kinect sensor.” 2015 IEEE 16th International Conference on Communication Technology (ICCT). IEEE, 2015.
4) Pullen, Paula, and William Seffens. ”Machine learning gesture analysis of yoga for exergame development.” IET Cyber-Physical Systems: Theory Applications 3.2 (2018): 106-110.
5) Trejo, Edwin W., and Peijiang Yuan. ”Recognition of Yoga poses through an interactive system with Kinect

device.” 2018 2nd International Conference on Robotics and Automation Sciences (ICRAS). IEEE, 2018.

6) Kodama, Tomoya, Tomoki Koyama, and Takeshi Saitoh. ”Kinect sensor-based sign language word recognition by multi-stream HMM.” 2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE). IEEE, 2017.
7) Calin, Alina Delia. ”Variation of pose and gesture recognition accuracy using two Kinect versions.” 2016 International Symposium on Innovations in Intelligent Systems and Applications (INISTA). IEEE, 2016.
8) University of Rochester Medical Center. (2020). Top 10 Most Common Health Issues [Online]. Available: https://www.urmc.rochester.edu/senior-health/common-issues/topten.aspx
9) Z. Cao, T. Simon, S. Wei and Y. Sheikh, ”Real-time Multi-person 2D Pose Estimation Using Part Affinity Fields,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 1302-1310.