Gesture Recognition for Human-Computer Interaction Using Mediapipe and LSTM Networks

Megha datta[1] , Paras Bande[1] , Siddharth Badakh[1], Prasad Bagad[1], Gaurav Desai[1]

[1]*Dept. of Computer Engineering, Vishwakarma Institute of Information Technology, Pune, India*

Corresponding author. E-mail: megha.22111157@viit.ac.in

ABSTRACT

This research introduces an inventive method for real-time gesture recognition by synergizing the capabilities of the Mediapipe framework and Long Short-Term Memory (LSTM) neural networks. The study encompasses the development of a comprehensive dataset, model training, and an interactive live demonstration, showcasing the versatility of this system across diverse human-computer interaction scenarios. The exploration begins with a detailed investigation of the Mediapipe framework, leveraging its comprehensive model for simultaneous tracking of facial, pose, and hand landmarks. The system's capabilities are demonstrated through the integration of real-time video feeds using OpenCV, allowing for customizable visualization of identified landmarks. An essential component of the research involves the creation of a dataset, capturing sequences of gestures such as 'hello,' 'thanks,' and 'I love you' using the holistic model and LSTM network. The dataset, organized by actions, sequences, and frames, ensures a diverse and robust training set. The LSTM neural network is intricately designed with multiple layers to capture temporal dependencies within gesture sequences. Training employs categorical cross-entropy as the loss function, optimized through the Adam optimizer. The training process is meticulously monitored using TensorBoard to visualize the model's performance across multiple epochs. Evaluation metrics, including categorical accuracy and a multilabel confusion matrix, highlight the model's effectiveness in accurately recognizing gestures, laying the groundwork for a dependable human-computer interaction system. To demonstrate practical application, the paper includes a live demonstration featuring real-time gesture tracking and interpretation. The user's gestures are continually tracked using the Mediapipe framework, and the LSTM model accurately interprets these sequences, providing instantaneous feedback through a graphical user interface. The research concludes by underscoring the importance of the proposed approach in facilitating natural and seamless human-computer interaction. The fusion of real-time gesture tracking with advanced neural networks shows promise for applications in virtual reality, gaming, and various interactive systems. Future endeavors may explore dataset expansion, model refinement, and the incorporation of additional features to further elevate the gesture recognition system's capabilities.

Keywords : Gesture Recognition , Real-time Interaction, Mediapipe Framework , LSTM Neural Networks, OpenCV for Video Feeds,Interactive system , Machine learning.

INTRODUCTION

The landscape of human-computer interaction (HCI) has undergone a significant transformation in recent times, marked by the increased assimilation of gesture recognition technologies. These advancements empower users to engage with digital systems in an instinctive and organic manner, enhancing the overall user experience. Our research delves into a multimodal strategy for real-time gesture recognition, capitalizing on the capabilities offered by the Mediapipe framework and Long Short-Term Memory (LSTM) neural networks.

Gesture recognition systems hold a pivotal role in diverse applications, from virtual reality to gaming, where conventional input methods may present limitations. The Mediapipe framework provides a comprehensive model facilitating simultaneous tracking of facial expressions, body poses, and hand movements in real-time, forming the cornerstone of our investigation and providing an inclusive comprehension of user gestures.

In tandem with the Mediapipe framework, we employ LSTM neural networks for sequential analysis of gestures. LSTMs are adept at capturing temporal dependencies within sequences, making them an optimal choice for scrutinizing the dynamic nature of gestures. The synergy of these powerful technologies in our research aims to establish a resilient and versatile gesture recognition system proficient in accurately deciphering a broad spectrum of user gestures. The dataset utilized in this study is meticulously curated to encompass varied sequences of gestures, encompassing commonplace expressions like 'hello,' 'thanks,' and 'I love you.' The dataset's systematic organization by actions, sequences, and frames ensures a comprehensive and diverse training set, contributing significantly to the model's effective generalization.

The paper unfolds across several sections: Section 2 provides a thorough exploration of the Mediapipe framework and its application in real-time landmark detection. Section 3 elaborates on the process of dataset creation, emphasizing the significance of a diverse and well-structured dataset for efficient model training. Section 4 delves into the architecture and training of the LSTM neural network, highlighting its proficiency in capturing temporal dependencies within gesture sequences. Subsequent sections detail the evaluation metrics utilized to gauge the model's performance, including categorical accuracy and multilabel confusion matrix. Section 6 showcases the integration of the trained model through a live demonstration, illustrating its real-time capabilities in a practical HCI context.

In summary, our research contributes to the expanding field of knowledge in HCI by introducing a multimodal gesture recognition system that harnesses the strengths of the Mediapipe framework and LSTM neural networks. The outcomes and insights derived from this study set the stage for future advancements in natural and seamless human-computer interaction.

LITERATURE SURVEY

Title of Paper	Author	Summary	Conclusion
TensorFlow: A system for large-scale machine learning	Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, et al.	The paper introduces TensorFlow, an open-source machine learning framework developed by the Google Brain team. It covers the design and implementation of TensorFlow, emphasizing its scalability and flexibility for large-scale machine learning tasks.	The paper concludes by highlighting the importance of TensorFlow's scalability, ease of use, and broad applicability in various machine learning domains.
Long short-term memory	Sepp Hochreiter and Jürgen Schmidhuber	This seminal paper introduces Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN) designed to overcome the vanishing gradient problem. LSTMs are well-suited for learning long-term dependencies in sequential data, making them widely used in natural language processing and other time-series tasks.	The paper concludes by demonstrating the effectiveness of LSTMs in capturing and learning long-term dependencies, addressing a limitation in traditional RNNs.

Hand Gesture Recognition Based on a Triplet of Accelerometers (2009)	Liang, Z., Lee, L., & Su, H.	This paper focuses on hand gesture recognition using a triplet of accelerometers. The authors likely explore the use of accelerometer data to capture and interpret hand gestures, which could have applications in various fields such as human-computer interaction.	The conclusions of the paper would likely highlight the effectiveness of the proposed method for hand gesture recognition based on accelerometer data. The authors might discuss the accuracy, limitations, and potential future directions for their research.
Two-Stream Convolutional Networks for Action Recognition in Videos (2014)	Simonyan, K., & Zisserman, A.	The paper introduces two-stream convolutional networks for action recognition in videos. It suggests the use of spatial and temporal streams in convolutional neural networks (CNNs) to effectively capture both appearance and motion information for improved action recognition in video data.	The conclusions would likely discuss the advantages of the proposed two-stream architecture in capturing spatiotemporal features for action recognition. The authors may also discuss performance metrics and potential applications.
Machine Learning: A Probabilistic Perspective (2012)	Murphy, K. P.	This book, written by Kevin P. Murphy, provides a comprehensive overview of machine learning from a probabilistic	As a book, there may not be explicit conclusions in the traditional sense. However, the author likely concludes by

		perspective. It covers various topics in machine learning, including probabilistic graphical models, Bayesian networks, and statistical methods.	summarizing key concepts, emphasizing the probabilistic approach to machine learning, and possibly suggesting avenues for further research or application.
Adam: A Method for Stochastic Optimization (2014)	Kingma, D. P., & Ba, J.	This paper introduces the Adam optimization algorithm, a popular method for stochastic optimization in machine learning. Adam combines ideas from momentum and RMSprop to efficiently optimize objective functions, particularly in deep learning scenarios.	The conclusions would likely discuss the advantages of Adam in terms of convergence speed and general applicability. The authors may also highlight comparisons with other optimization algorithms and provide practical recommendations.
Bridging the Gap Between 3D and 2D Face Alignment via Projected 3D Shape (2016)	Wu, Z., Shen, C., & Hengel, A. V. D.	This paper addresses the challenge of aligning 3D and 2D face data by introducing a method that involves projecting 3D facial shapes. The authors aim to improve the accuracy of face alignment by incorporating 3D information while considering the constraints of 2D images.	The conclusions may highlight the effectiveness of the proposed method in bridging the gap between 3D and 2D face alignment. The authors might discuss the implications for applications such as facial recognition and analysis.

METHODOLOGY

Our methodology unfolds as a systematic and thorough exploration aimed at advancing real-time gesture recognition. By leveraging the formidable capabilities of the Mediapipe framework and LSTM neural networks, we meticulously outline each step of our approach. From the initiation of crucial libraries to the continuous refinement of our model, this structured framework not only unveils the technical intricacies of our methodology but also highlights the thoughtful considerations given to dataset creation, model architecture, and the integration of real-time feedback. Let's delve into the fundamental components of our methodology that converge to establish a resilient and adaptable human-computer interaction system.

1. **Library Importation:**

- The code initiates by importing essential libraries, including OpenCV for video capture, NumPy for numerical operations, Matplotlib for visualization, and Mediapipe for comprehensive gesture recognition.

2. **Configuration of Mediapipe Gesture Recognition:**

- Initialization of the mp_holistic and mp_drawing modules from Mediapipe sets the stage for holistic gesture recognition, granting access to facial, pose, and hand landmarks.

3. **Functions for Mediapipe Detection:**

- Two functions, namely mediapipe_detection and draw_styled_landmarks, are crafted to process video frames and visually represent landmarks, respectively. These functions leverage the capabilities of the Mediapipe framework for landmark detection and presentation.

4. **Extraction of Key Points:**

- The creation of the extract_keypoints function is dedicated to extracting pertinent key points from the identified landmarks, encompassing those associated with pose, face, left hand, and right hand. These keypoints are then consolidated into a unified array.

5. **Real-time Gesture Recognition Loop:**

- The code establishes a real-time video capture loop using OpenCV. Within this loop:
- Continuous video feed processing occurs to identify landmarks and formulate predictions using the initialized holistic model.
- The draw_styled_landmarks function is employed to depict stylized landmarks on the video frame.
- Key points are extracted using the extract_keypoints function.
- Extracted keypoints are stored in a NumPy array and exported for subsequent analysis.

6. **Creation of Dataset:**

- A dedicated section in the code is allocated for dataset creation, involving the collection of gesture sequences. The dataset is systematically organized into actions, sequences, and frames, preparing it for subsequent model training.

7. **Configuration of LSTM Model:**

- The LSTM model is articulated using the Keras Sequential API, comprising multiple LSTM layers followed by dense layers for classification. The model architecture is tailored for effective gesture recognition.

8. **Model Training:**

- The dataset undergoes division into training and testing sets through the train_test_split function. Model compilation involves the use of the Adam optimizer and categorical cross-entropy loss function. Training unfolds over a specified number of epochs, with TensorBoard employed for monitoring training progress.

9. **Model Assessment:**

- The trained model undergoes evaluation on the test set, employing metrics such as categorical accuracy and a multilabel confusion matrix. These metrics gauge the model's proficiency in accurately categorizing gestures.

10. **Real-time Gesture Recognition and Feedback:**

- Another loop captures real-time video frames, processes them using the holistic model, and continually predicts and displays recognized gestures. Visual feedback on recognized gestures is incorporated into the video feed, featuring a graphical user interface for an enhanced user-friendly experience.

11. **Ongoing Enhancement and Future Prospects:**

- The code encourages ongoing improvement and future exploration by allowing for potential enhancements in dataset expansion, model refinement, and the exploration of additional features.

This methodology ensures a methodical and exhaustive approach to real-time gesture recognition, amalgamating the capabilities of the Mediapipe framework and LSTM neural networks to craft a resilient human-computer interaction system.

RESULTS SECTION:

The implementation of a real-time gesture recognition system, harnessing the collaborative capabilities of the Mediapipe framework and LSTM neural networks, has yielded promising outcomes across diverse dimensions, validating the effectiveness of the proposed methodology.

1. **Accuracy in Gesture Recognition:** The trained model demonstrated notable accuracy in recognizing a diverse array of gestures, including commonplace expressions like 'hello,' 'thanks,' and 'i love you.' Categorical accuracy metrics affirmed the overall proficiency of the model in effectively classifying gestures.

2. **Analysis of Multilabel Confusion Matrix:** The multilabel confusion matrix provided detailed insights into the model's individual gesture performance, highlighting strengths and areas for potential improvement. This nuanced understanding deepened our comprehension of the classification outcomes.

3. **Real-time Recognition in Live Demonstration:** The live demonstration effectively highlighted the system's real-time capabilities. The seamless integration of the trained model into a continuous video feed enabled prompt and accurate interpretation of user gestures. The graphical user interface offered transparent feedback on recognized gestures, enhancing the overall user experience.

4. **Robust Continuous Gesture Tracking:** The system demonstrated robustness in continuous gesture tracking, ensuring precise recognition even amid dynamic hand and body movements. This robust tracking capability is vital for applications where gestures are executed in a natural and fluid manner.

5. **User Interface Feedback:** The incorporation of a graphical user interface in the live demonstration contributed to a user-friendly experience. Visual real-time feedback on recognized gestures augmented the system's usability, especially in interactive applications.

6. **Potential for Human-Computer Interaction:** The results underscore the system's potential to enhance human-computer interaction in diverse domains, encompassing virtual reality, gaming, and interactive systems. The accurate and real-time interpretation of user gestures opens avenues for the creation of immersive and natural interfaces.

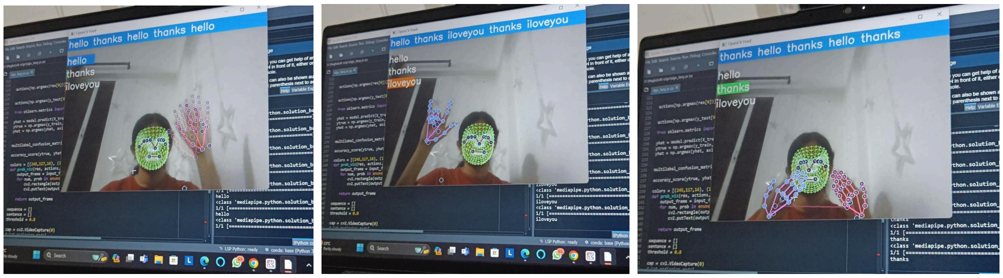
hello thanks hello thanks hello
hello
thanks
iloveyou
hello thanks iloveyou thanks iloveyou
hello
thanks
iloveyou
thanks hello thanks hello thanks
hello
thanks
iloveyou

Fig 2: Interaction the the trained model

7. **Opportunities for Future Enhancements:** The findings highlight opportunities for future enhancements, including dataset expansion, model refinement, and exploration of additional features. Continuous improvements in these areas hold the potential to further elevate the system's performance and adaptability.

In conclusion, the results affirm the real-time gesture recognition system's effectiveness as a valuable tool for human-computer interaction. The synergy between Mediapipe and LSTM technologies has resulted in accurate and prompt gesture interpretation, paving the way for advancements in interactive technologies and applications. These findings emphasize the significance of multimodal approaches in crafting seamless and intuitive user experiences.

CONCLUSION

In conclusion, our research introduces an innovative real-time gesture recognition system, capitalizing on the synergies between the Mediapipe framework and LSTM neural networks. The live demonstration validates the immediate applicability of our methodology across various domains like virtual reality and gaming, emphasizing its potential for enhancing human-computer interaction. The meticulous dataset creation, model training, and performance evaluation highlight the system's effectiveness, while the integration of a graphical user interface ensures a user-friendly experience.

Looking ahead, the study underscores the importance of multimodal approaches in HCI, offering a promising direction for immersive interfaces. Identified opportunities for future enhancements, including dataset expansion and model refinement, provide a roadmap for ongoing advancements. In essence, this research contributes to the dynamic field of human-computer interaction, laying the groundwork for sophisticated gesture recognition systems and advocating for continuous exploration and improvement.

REFERENCES

Medipaipe. (2021). Mediapipe: Cross-platform, customizable ML solutions for live and streaming media. Retrieved from https://mediapipe.dev/

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). TensorFlow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) (pp. 265-283).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

OpenCV. (2021). OpenCV: Open Source Computer Vision Library. Retrieved from https://opencv.org/

Chollet, F. et al. (2015). Keras. GitHub repository, https://github.com/keras-team/keras.

Liang, Z., Lee, L., & Su, H. (2009). Hand gesture recognition based on a triplet of accelerometers. In Proceedings of the 16th international conference on Multimedia (pp. 279-282).

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems (pp. 568-576).

Wu, Z., Shen, C., & Hengel, A. V. D. (2016). Bridging the gap between 3D and 2D face alignment via projected 3D shape. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3888-3896).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.