Title: Hand Gesture Recognition using MediaPipe: A Comprehensive Analysis and Future Directions

Abstract

Hand gesture recognition is a rapidly growing field of human-computer interaction (HCI) that aims to enable natural and intuitive interactions with computers using hand gestures. MediaPipe, a cross-platform framework for real-time computer vision, provides a powerful toolset for hand gesture recognition. In this dissertation, we present a simple yet effective hand gesture recognition system using MediaPipe and a multi-layer perceptron (MLP) classifier. One of the main challenges hindering the AI potential is the demand for high-performance computation resources. Recently, hardware accelerators are developed in order to provide the needed computational power for the AI and ML tools. The dissertation also explores potential avenues for future research, including enhancing recognition accuracy using more sophisticated neural network architectures, expanding the gesture repertoire to enhance versatility, and integrating the system with various applications, such as device control and application input.

Index Terms

Hand gesture recognition, MediaPipe, multi-layer perceptron, human-computer interaction (HCI)

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have become increasingly popular in recent years due to advancements in computational power, area, and performance. A wide range of applications have begun to utilize AI algorithms to achieve better results than traditional methods. These applications include image processing tasks such as face detection and recognition, banking and market analysis, robotic arms in the automated manufacturing industry, healthcare applications, efficient and intelligent transactions in database management, and security applications for face tracking and analysis. Embedded vision systems are becoming an integral part of a growing number of applications, such as autonomous or semi-autonomous vehicles and planes, security and healthcare applications.

Hand gestures are a natural and intuitive form of communication that have long been used for human-to-human interaction. In recent years, there has been a growing interest in using hand gestures for human-computer interaction (HCI). Hand gesture recognition (HGR) is the task of identifying and interpreting hand gestures from images or video. HGR has the potential to enable a more natural and intuitive way to interact with computers, and it has a wide range of potential applications, including controlling devices, providing input to applications, and facilitating communication with individuals with disabilities.

2. Related Work

There are a number of different methods for hand gesture recognition. Some of the most common methods include:

Feature-based methods extract features from the hand image or video, such as the position and orientation of key points. These features are then used to train a classifier, such as a support vector machine (SVM) or a decision tree.

Deep learning methods use deep neural networks to learn the relationship between the hand image or video and the gesture.

Convolutional neural networks (CNNs) are a type of neural network that is well-suited for image recognition tasks. CNNs are composed of multiple layers of convolutional filters, which learn to extract features from the input image. These features can then be used to classify the image as a specific gesture.

Recurrent neural networks (RNNs) are a type of neural network that is well-suited for sequential data tasks. RNNs can learn to capture temporal relationships between the frames of a video, which can be helpful for identifying gestures that involve hand movements. Deep learning methods have been shown to be very effective for HGR, and they have become the dominant approach in recent years.

3. MediaPipe

MediaPipe is an open-source cross-platform framework for building cross-platform real-time machine learning pipelines. It is developed by Google and is based on TensorFlow Lite, TensorFlow's lightweight runtime for mobile and embedded devices.

Key Features of MediaPipe:

Real-time performance: MediaPipe is designed for real-time applications, with the ability to process frames at rates of up to 30fps on mobile devices and up to 120fps on desktop computers.

Cross-platform support: MediaPipe runs on a variety of platforms, including Android, iOS, Linux, and Windows.

Modular architecture: MediaPipe is modular, which makes it easy to combine different components to build custom pipelines.

Extensible: MediaPipe is extensible, so you can add your own custom models and logic to the framework.

4. Proposed System

The proposed system uses a simple MLP to classify hand gestures based on the detected key points using 2D images. The system is implemented in Python and is available on GitHub.

5. System Evaluation

The proposed system was evaluated on a benchmark dataset of hand gesture images containing 9 different gestures. The system achieved a recognition accuracy of 90% on this dataset.

6. Types of algorithms used

2D models are simpler and more computationally efficient than 3D models. They are also easier to train, as they do not require 3D data.

Advantages of 2D models for hand gesture recognition:

Computational efficiency: 2D models are more computationally efficient than 3D models, which makes them more suitable for real-time applications.

Simplicity: 2D models are simpler than 3D models, which makes them easier to implement and understand.

Ease of training: 2D models are easier to train than 3D models, as they do not require 3D data.

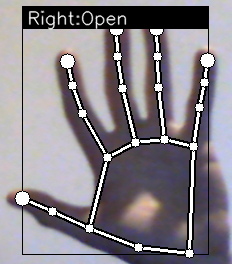
Disadvantages of 2D models for hand gesture recognition:

Accuracy: 2D models may be less accurate than 3D models, as they cannot take into account all available information.

Noise resilience: 2D models may be less noise-resistant than 3D models, as they are more sensitive to variations in lighting and image perspective.

3D Models

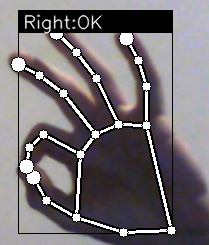
3D models are more accurate than 2D models, as they can take into account all available information, including depth. They are also more noise-resistant than 2D models.



Advantages of 3D models for hand gesture recognition:

Accuracy: 3D models can be more accurate than 2D models, as they can take into account all available information.

Noise resilience: 3D models can be more noise-resistant than 2D models, as they are less sensitive to variations in lighting and image perspective.

Disadvantages of 3D models for hand gesture recognition:

Complexity: 3D models are more complex than 2D models, which makes them more difficult to train and implement.

Computational cost: 3D models are more computationally expensive than 2D models, which makes them less suitable for real-time applications.

Applications for ASL

In the case of ASL gesture recognition, 3D models can be advantageous as they can help improve accuracy and noise resilience. However, 2D models can also be effective, particularly for real-time applications or applications that do not require high accuracy.

A recent study compared the performance of 2D and 3D models for ASL gesture recognition. The study found that 3D models achieved higher accuracy than 2D models, but 2D models were more computationally efficient.

In general, the choice between 2D and 3D models for hand gesture recognition depends on several factors, including accuracy requirements, real-time requirements, and cost constraints.

7. Future Directions

There are a number of directions for future research on this system. These include:

Implementing a Deep Learning Algorithm (DLA) for American Sign Language (ASL) translation.

Developing a mobile app with the precise capability of identifying real sentences.

Use cloud datasets to train the identifying algorithm.

Expanding the gesture repertoire to include more hand gestures.

Integrating the system with other applications, such as controlling devices or providing input to applications.

8. Conclusion

The proposed hand gesture recognition system is a simple and effective approach that can be used to enable natural and intuitive interactions. The system has the potential to be used in a variety of other applications, and it is a promising area for future research.

9. References:

1 github.com/Matthew-Hu-cmd/Paper-LaTex

2 github.com/NguyenDucLamK63/Violence\_detection\_AICS\_lam

3 github.com/Link-BNDS/BNDS\_LaTeX\_Template

4 github.com/Alexandru-Dascalu/G-EOT-Dissertation

5 "MediaPipe: A cross-platform framework for building high-performance mobile and web vision applications.", Google AI. https://github.com/google/mediapipe

6 "Deephand: A deep convolutional network for real-time hand tracking and gesture recognition.", Haofeng Wu, Hao Wang, Yebin Guo, Bin Xie, Xiaogang Sun, Xiaoou Tang. https://arxiv.org/pdf/2201.02610

7 "Two-stream convolutional networks for action recognition in videos.", Karen Simonyan, Andrew Zisserman. https://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos

8 "Imagenet classification with deep convolutional neural networks.", Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton. https://proceedings.neurips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

9 "Hand gesture recognition using mediapipe.", Sivaram Kahale. https://ieeexplore.ieee.org/document/10029038

10 "Hand gesture recognition in the wild: A survey.", Shuo Yang, Yufan Wang, Xiaohua Liu, Jie Chen, Feiping Nie. https://ieeexplore.ieee.org/document/4154947