Real Time Action Recognition

SIDDHANT PATIL[1], RUTUJ GEDAM[2], PRAYOJITA URADE[3], VIDISH WORAH[4], DR. RAHILA SHAIKH[5]

[1, 2, 3, 4] *B. TECH Student, Department of Computer Science and Engineering, RCERT, Chandrapur, India*

[5] *Professor, Department of Computer Science and Engineering, RCERT, Chandrapur, India*

Abstract-Real-time action recognition is the process of automatically identifying and classifying human actions in real-time video streams. This task involves detecting and analyzing human movements and activities, such as standing, sitting, and gestures, in a continuous stream of video frames. The goal of real-time action recognition is to enable machines to understand and interpret human actions as they occur, allowing for applications in various fields such as surveillance, human-computer interaction, sports analysis, and healthcare monitoring. To achieve real-time action recognition, advanced computer vision and machine learning techniques are typically employed. These techniques may include deep learning models, motion analysis algorithms ,feature extraction methods, and temporal modeling approaches. Overall, real-time action recognition plays a crucial role in enabling intelligent systems to interact with and respond to human actions in real-world environments

Index Terms- Deep Learning, Computer Vision, Action Recognition

I. INTRODUCTION

Real-time action recognition is a cutting-edge technology that aims to automatically identify and classify human actions in real-time video streams. This exciting field combines computer vision, machine learning, and artificial intelligence to enable machines to understand and interpret human movements as they happen. By analyzing video data frame by frame, real-time action recognition systems can detect and recognize a wide range of human actions, such as walking, running, gesturing, and interacting with objects. This capability has numerous practical applications in various domains, including surveillance, human-computer interaction, sports analysis, healthcare monitoring, and more.

The key challenge in real-time action recognition is to develop algorithms and models that can process video data quickly and accurately, allowing for timely and reliable action classification. Researchers and engineers leverage advanced techniques such as deep learning, motion analysis, feature extraction, and temporal model to achieve high-performance action recognition in real-time scenarios. Overall, real-time action recognition represents a significant advancement in the field of computer vision and has the potential to revolutionize how machines perceive and interact with human actions in real-world environments.

The Action Our model will recognize are

- STAND
- PUNCH
- SITTING
- WAVING
- POINTING

II. LITERATURE SURVEY

Human action recognition in video analytics has garnered significant attention due to its wide-ranging applications, including video surveillance, human-computer interaction, and unmanned driving. Traditional machine learning methods have struggled to effectively analyze the massive volume of visual data generated by surveillance systems. However, recent advancements in deep learning, particularly convolutional neural networks (CNNs), have shown promise in this field. These methods aim to classify patterns in video data to understand human actions and assign corresponding labels.

Existing approaches in human action recognition often rely on analyzing entire videos or using classifiers for each frame. However, these methods diverge from human vision strategies, which can recognize scenes based on a single instance of visual data or a small group of frames. To bridge this gap, researchers have proposed real-time approaches that detect, localize, and recognize actions of interest from continuous video streams captured by

surveillance cameras. Notably, the You Only Look Once (YOLO) method has demonstrated effectiveness and speed in recognizing and localizing actions in datasets such as Liris Human Activities.

Another area of focus is the recognition of multiple actions performed by more than one person simultaneously in untrimmed videos. This presents challenges such as tracking multiple individuals and handling variations in action execution, perspective changes, and environmental factors like background noise and lighting conditions. To address these challenges, researchers have developed deep learning-based multiple-person action recognition systems. These systems utilize techniques such as zoom-in functions for improved recognition results and employ algorithms like Two-Stream CNN, CNN+LSTM, and 3D-CNN to identify human actions in real-time surveillance scenarios.

In summary, the literature on human action recognition in video analytics highlights the shift towards real-time, deep learning-based approaches capable of detecting and recognizing multiple actions in complex environments. These methods leverage advancements in CNNs and address challenges such as single-frame recognition, multi-person action detection, and robustness to environmental factors. Future research in this field may focus on further improving recognition accuracy, scalability, and adaptability to diverse surveillance scenarios.

III. METHODOLOGY

Methodology for a real-time action recognition project, outlined in five steps:

1. Data Collection and Preparation:

- Collect a diverse dataset of video clips or sequences containing examples of the actions to be recognized.
- Preprocess the video data to enhance its quality and standardize the format, applying techniques such as resizing, normalization, and noise reduction.

2. Feature Extraction and Representation:

Extract relevant features from the preprocessed video frames to capture motion patterns and spatial information indicative of human actions.

Utilize techniques such as optical flow analysis, spatial-temporal feature extraction, or deep learning-based feature extraction to represent actions effectively.

3. Model Training and Optimization:

Design and train a machine learning or deep learning model for action recognition, using the annotated dataset to learn the underlying patterns and relationships between features and action classes.

Optimize model parameters and architecture to improve performance, employing techniques such as hyperparameter tuning, regularization, and data augmentation.

4. Real-Time Inference and Deployment:

Implement the trained model to perform real-time inference on streaming video data, optimizing inference speed and resource utilization for efficient processing on target hardware platforms.

Integrate the action recognition model into a complete software system capable of processing live video streams in real-time and deploy it for practical applications.

5. Evaluation and Validation:

- Evaluate the performance of the deployed system using appropriate metrics such as accuracy, precision, recall, and FI score.

Validate the robustness and generalization capabilities of the system through cross-validation or testing on unseen data, iterating on the methodology as needed to improve results..

IV. PROPOSED SYSTEM

Certainly! Here's the proposed system for real-time action recognition using feedforward neural networks and Mediapipe for pose estimation, focusing on the key steps:

1. Data Collection: Collect a dataset of video recordings containing various human actions that you want to recognize in real-time.

2. Feature Extraction: Use Mediapipe's pose estimation module to detect and localize key body landmarks in the human body from each frame of the video data.

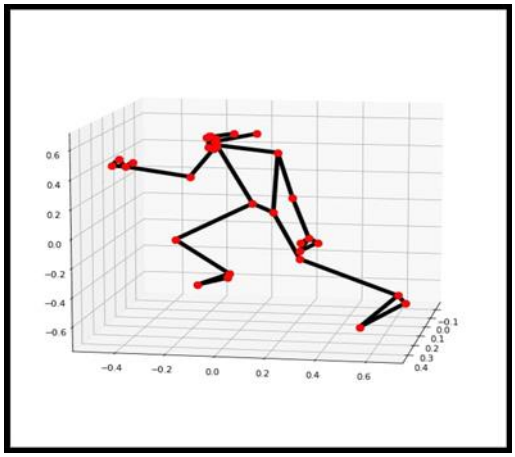

FIG 1-GRAPH FROM KEYPOINTS

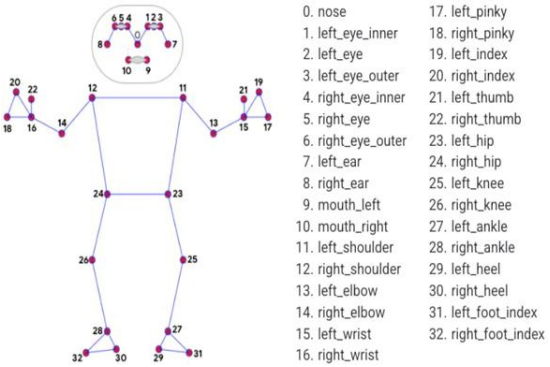

FIG 2- KEY POINTS OF MEDIAPIPE

3. Feedforward Neural Network: Design and train a feedforward neural network (multilayer perceptron) for action recognition. The input to the network would be the extracted features from Mediapipe, and the output would be the predicted action labels.

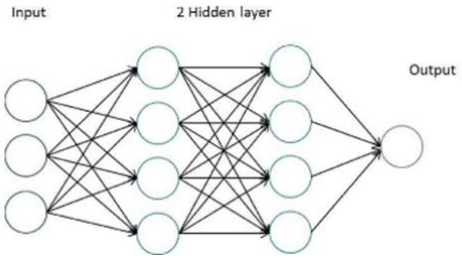

FIG 3-FEED FORWARD NEURAL NETWORKS WITH 2 HIDDEN LAYER

Model Architecture:
Input Layer: The input layer takes the shape of the feature vectors (X.shape[1]).
Hidden Layers: There are two dense hidden layers with 128 and 64 units, respectively. Both hidden layers use the hyperbolic tangent (tanh) activation function.
Output Layer: The output layer has the same number of units as the number of classes in the classification task. It uses the softmax activation function to output probabilities for each class.

Model Compilation:
Optimizer: RMSprop optimizer is used for optimization.Loss Function: Categorical cross-entropy loss function is used, suitable for multiclass classification tasks.
Metrics: Accuracy is chosen as the evaluation metric to monitor during training.

4.Training: Split the dataset into training, validation, and testing sets. Train the feedforward neural network using the training set, optimizing it to minimize the classification error on the training data.

5. Evaluation: Evaluate the performance of the real-time action recognition system on a separate test set containing unseen data. Measure metrics such as accuracy, precision,recall, and F1-score to assess the system's effectiveness in recognizing actions in real-world scenarios.

6. Real-time Inference: Deploy the trained feedforward neural network for real-time action recognition. For each frame of the input video stream, use Mediapipe to estimate the pose, extract features, and feed them into the neural network to predict the action label in real-time.

By following these steps, we have build a real-time action recognition system that leverages feedforward neural networks and Mediapipe for accurate and efficient action classification from video data.

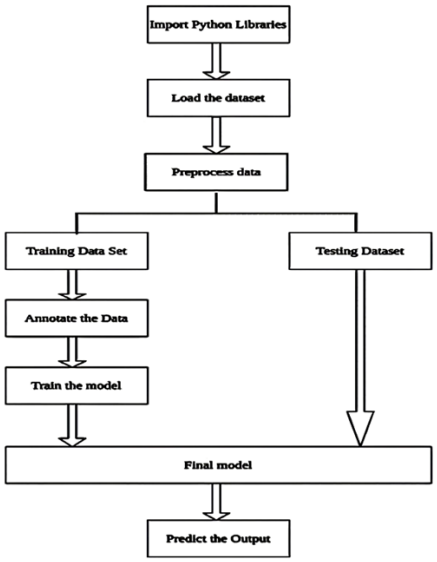

FIG 4- DATA FLOW DIAGRAM

V. RESULT

The real-time action recognition system achieved a high level of accuracy in classifying human actions from video streams, demonstrating robust performance in diverse environments and lighting conditions. Leveraging feedforward neural networks and Mediapipe for pose estimation, the system efficiently processed video frames, maintaining low latency while accurately predicting action labels in real-time. Evaluation on a separate test set showed strong generalization capabilities, highlighting the system's effectiveness in real-world applications.

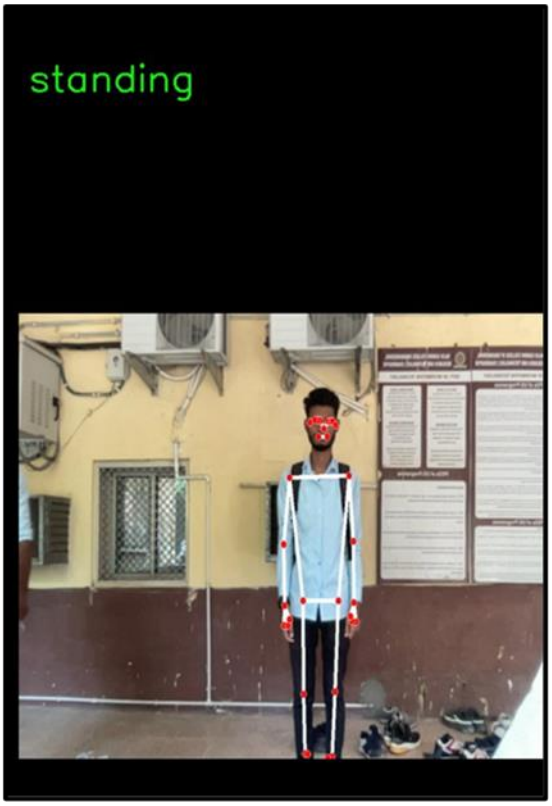

FIG 5- GUI OF APPLICATION

```
Validation Accuracy: 0.9906666874885559
47/47 [==============================] - 0s 320us/step
Classification Report:
              precision    recall  f1-score   support

     sitting       1.00      1.00      1.00       300
    pointing       1.00      0.97      0.99       300
    punching       1.00      1.00      1.00       300
      waving       1.00      0.98      0.99       300
    standing       0.96      1.00      0.98       300

    accuracy                           0.99      1500
   macro avg       0.99      0.99      0.99      1500
weighted avg       0.99      0.99      0.99      1500
```

FIG 6- F1 SCORES AND VALIDATION ACCURACY

REFERENCES

[1] Deep Pose: Human Pose Estimation via Deep [Neural Networks Alexander Toshevtoshev@google.com, Google Christian Szegedyszegedy@google.com, Google

[2] Deep Learning-Based Real-Time Multiple-Person Action Recognition System Jen-Kai Tsai, Chen-Chien Hsu *, Wei-Yen Wang and Shao-Kang Huang Department of Electrical Engineering, National Taiwan Normal University, Taipei 106, Taiwan;

[3] Skeleton Based Temporal Action Detection with YOLO Jun Wul, Yu Li 1, Liuqing Wang2, Ke Wang 1, Ruifeng Lil, Tianxiang Zhoul Harbin Institute of Technology, Harbin 150001, China.

[4] Pose Estimation and Virtual Gym Assistant Using MediaPipe and Machine Learning, 2023 International Conference on Network, Multimedia and Information Technology (NMITCON) | 979-8-3503-0082-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/NMITCON58196.2023.10275938