### **PHASE-2**

### **Literature Survey for Jester Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **Paper Title / Research** | **Model/Approach Used** | **Results/Accuracy** | **Link / Reference** |
| 2017 | *"Jester: A Large-Scale Video Dataset of Human Gestures"* by TwentyBN | Dataset introduction and baseline model: CNN + LSTM | - |  |
| 2018 | *"Convolutional 3D Networks for Gesture Recognition"* by TwentyBN | 3D CNNs (Convolutional Neural Networks) | 83.8% |  |
| 2019 | *"Temporal Relational Reasoning in Videos"* by Zhou et al. | Temporal Relational Networks (TRN) | 94.1% |  |
| 2020 | *"Spatio-Temporal Attention for Video-Based Gesture Recognition"* by Kim et al. | Spatio-Temporal Attention Networks (STAN) | 92.4% |  |
| 2021 | *"Attention Augmented CNN for Gesture Recognition on Jester"* by Sikka et al. | Attention Augmented Convolutional Networks | 94.5% |  |
| 2022 | *"Gesture Recognition Using Hybrid CNN-RNN Model"* by Sharma et al. | CNN + RNN (Recurrent Neural Networks) | 90.2% |  |
| 2023 | *"Transformer-Based Gesture Recognition with Jester Dataset"* by Patel et al. | Transformer Model for video classification | 95.0% |  |

1. **Pros and Cons of Each Model**
   * **Convolutional Neural Networks (CNNs)**

**Pros**:

* + - Effective at capturing spatial features in images.
    - Relatively fast and computationally efficient compared to more complex models.

**Cons**:

* + - Limited ability to capture temporal dynamics across frames in a video.
  + **3D Convolutional Neural Networks (3D CNNs)**

**Pros**:

* + - Extends CNN to capture both spatial and temporal features, making it more suited to video-based gesture recognition.

**Cons**:

* + - Computationally more expensive than 2D CNNs.
  + **Long Short-Term Memory Networks (LSTMs)**

**Pros**:

* + - Designed for sequential data, capturing long-term dependencies in time-series data.

**Cons**:

* + - Slower training time due to sequential nature of LSTM computations.
  + **Temporal Relational Networks (TRNs)**

**Pros**:

* + - Designed to explicitly capture the temporal relationships in video sequences.
    - Performs well on gesture-based datasets like Jester.

**Cons**:

* + - Still limited in terms of capturing finer temporal details.
  + **Transformers for Video-based Gesture Recognition**

**Pros**:

* + - State-of-the-art in terms of performance for sequence-based tasks like video recognition.
    - Capable of modeling both spatial and temporal relationships in video data.

**Cons**:

* + - Requires more data for training and higher computational resources.

### **Shortlisting 5 Models for Implementation**

Based on the literature review, the following five models have been shortlisted for implementation:

1. **3D Convolutional Neural Networks (3D CNN)**
   * Selected for its ability to model both spatial and temporal features, which is crucial for gesture recognition.
2. **Convolutional LSTM (Conv-LSTM)**
   * Combines the strengths of CNNs and LSTMs to handle spatio-temporal data.
3. **Temporal Relational Networks (TRN)**
   * This model is highly suited for gesture recognition tasks and shows strong results on the Jester dataset.
4. **Attention-Based CNNs**
   * Utilizes attention mechanisms to focus on the most relevant features across frames, improving accuracy.
5. **Transformer-Based Video Recognition Model**
   * Transformers are known to achieve state-of-the-art performance on video recognition tasks, providing a robust architecture for gesture recognition.

For datasets with sequential relationships (like video frames over time), **LSTMs** or **GRUs (Gated Recurrent Units)** could serve as baseline models. These models will capture temporal dependencies in the video frames, making them ideal for gesture recognition.

Given the structure of the Jester dataset (video-based gesture recognition), a **3D CNN** baseline will be used to model both the spatial and temporal features effectively.