ACTION RECOGNITION FROM REAL-TIME VIDEOS

ABSTRACT

In this study, we propose a method for human action recognition in the UCF50 dataset by combining Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs). The UCF50 dataset poses challenges like lighting variations and background clutter, making it ideal for real-world scenario analysis. Our approach utilizes CNNs to extract spatial features from individual frames and LSTM networks to capture temporal dependencies across frames. Employing a two-stream architecture, one stream processes optical flow images for motion information while the other handles RGB frames for appearance information. We enhance feature extraction by fine-tuning pre-trained CNN models on large-scale image datasets and train the LSTM network jointly with the CNN stream for end-to-end learning of spatiotemporal representations. Experimental results showcase the method's competitive performance against state-of-the-art approaches, demonstrating robustness to action types, backgrounds, and viewing angles, thus highlighting its potential for diverse applications in human action recognition.

INTRODUCTION

Human action recognition from video data plays a pivotal role in various domains, including surveillance, human-computer interaction, and sports analysis. The ability to automatically understand and classify human actions has significant implications for enhancing security measures, developing intelligent systems, and understanding human behavior. With the proliferation of video data from diverse sources such as surveillance cameras, social media platforms, and wearable devices, the demand for accurate and efficient action recognition algorithms has grown substantially.

The UCF50 dataset stands as a benchmark for evaluating action recognition algorithms, comprising a wide range of human actions captured in realistic settings. However, recognizing actions in this dataset poses several challenges, including variations in lighting conditions, background clutter, and viewpoint changes. Addressing these challenges requires the development of sophisticated algorithms capable of capturing both spatial and temporal information inherent in video sequences.

In recent years, deep learning approaches, particularly Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have emerged as powerful tools for analyzing visual data, including videos. CNNs excel in extracting spatial features by analyzing individual frames, while LSTMs are effective in capturing temporal dependencies across frames. Leveraging the complementary strengths of CNNs and LSTMs has shown promise in improving the accuracy of action recognition systems.

➢ Problem Definition

This study focuses on accurately classifying human actions in the UCF50 dataset by leveraging CNNs and LSTM networks to capture spatial and temporal information, overcoming challenges like lighting variations and background clutter for robust recognition.

ALGORITHM

1.LSTM (Long Short Term Memory) :

1. **Input Sequence:** LSTM receives input data in the form of a sequence, such as a sequence of words in natural language processing or a sequence of frames in video analysis.

2. **Forget Gate:** The LSTM unit begins by determining which information from the previous cell state should be discarded or forgotten based on the current input and the previous hidden state.

3. **Input Gate:** It then decides which new information from the current input should be added to the cell state.

4. **Cell State Update:** The forget gate output and the input gate output are combined to update the cell state, allowing the LSTM to retain relevant information over long sequences.

5. **Output Gate:** Finally, the LSTM determines which parts of the cell state should be exposed as the output, which is then passed on to the next LSTM unit or the final output layer for prediction.

6. **Hidden State:** At each time step, the LSTM unit maintains a hidden state, which serves as a memory of past information and influences future predictions.

2.CNN (Convolutional Neural Networks) :

- **Input Image:** The algorithm begins with an input image or a series of images in the case of video data.

- **Convolution Operation:** CNN applies a series of convolutional filters (kernels) to the input image(s) to extract various features. Each filter detects specific patterns or features, such as edges or textures.

- **ReLU Activation:** After each convolution operation, a Rectified Linear Unit (ReLU) activation function is applied element-wise to introduce non-linearity, enhancing the model's ability to learn complex patterns.

- **Pooling:** Pooling layers (e.g., max pooling) are employed to downsample the feature maps, reducing computational complexity and extracting the most relevant features.

- **Flattening:** The feature maps are then flattened into a vector, preparing them for input into a fully connected layer.

- **Fully Connected Layers:** The flattened features are fed into one or more fully connected layers, which perform classification by learning complex relationships between features and outputting probabilities for each class.

- **Softmax Activation:** A softmax activation function is applied to the output layer to convert the raw scores into probability distributions over the classes, enabling the model to make predictions.

Convolutional Neural Networks (CNNs) for Feature Extraction

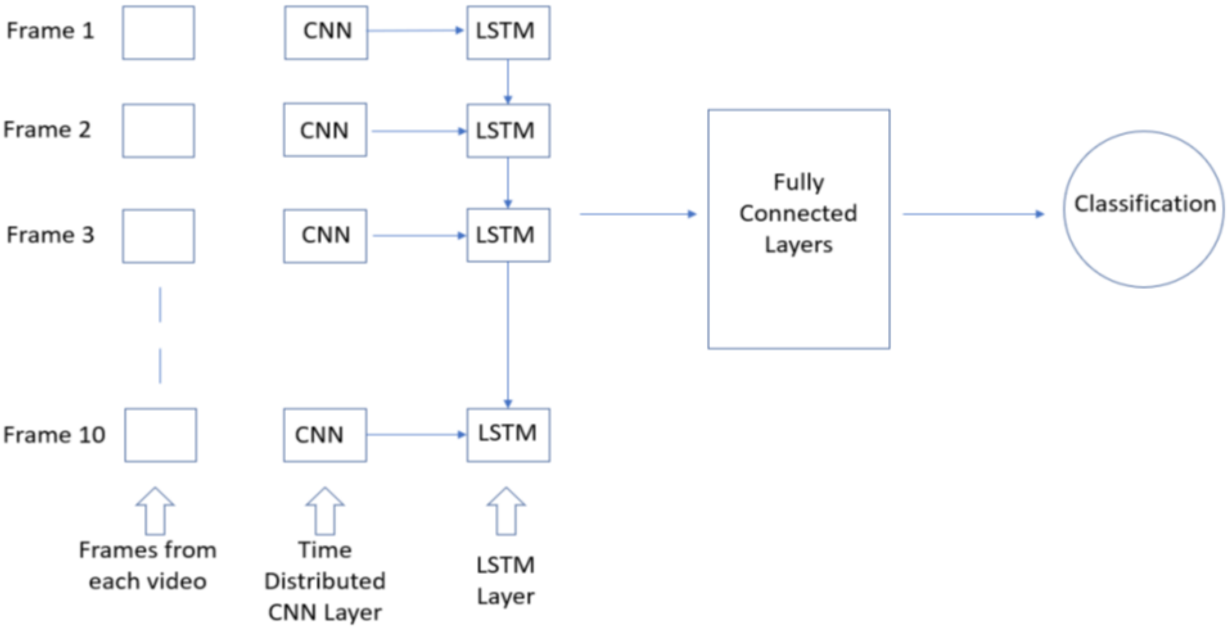
CNNs are the backbone of our action recognition model. In this section, we'll explore CNNs' role in extracting spatial features from video frames. We'll discuss popular CNN architectures, such as VGG16 and ResNet, which serve as powerful feature extractors. Additionally, we'll introduce our custom CNN architecture tailored for the UCF50 dataset.

Long Short-Term Memory (LSTM) Networks for Temporal Modeling

Action recognition isn't just about spatial features; it's also about temporal dependencies. LSTM networks come to the rescue for capturing these temporal relationships. We'll delve into the theory behind LSTMs and their unique ability to model sequences effectively. You'll gain insights into the architecture of our LSTM-based model designed for UCF101. We'll address the challenge of handling video sequences as input data.

Model Integration: CNN + LSTM

The magic happens when we combine CNN and LSTM layers to form our action recognition model. We'll explore the concept of 3D convolution, which blends spatial and temporal features seamlessly. We'll present the architecture of our integrated CNN-LSTM model, including input shapes and layer connections.



PERFORMANCE METRICS

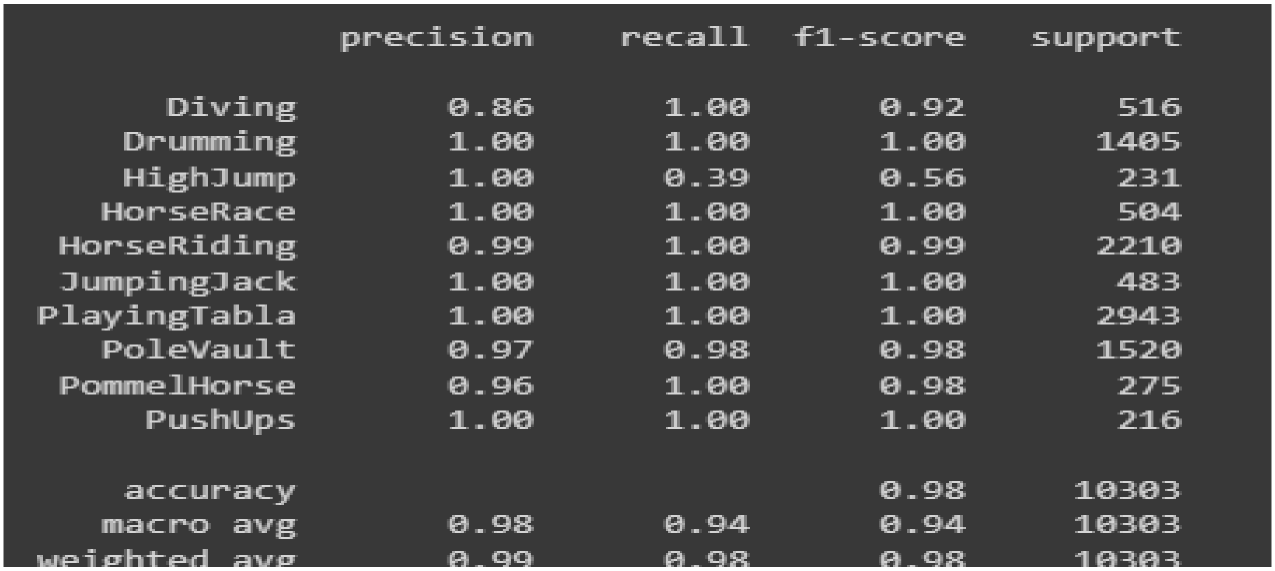
In evaluating the performance of human action recognition systems, several metrics are commonly used to assess their accuracy, robustness, and efficiency. Some of the key performance metrics include:

Accuracy:

$$accuracy = \frac{TR}{TR + FR} \qquad (7)$$

The accuracy of the recognition method was calculated
the number of samples respectively with successful and failed recognition. Furthermore, to demonstrate the advantage of the proposed action recognition method for different poses, the two methods based on different feature extraction methods as described were compared. On this basis, the confusion matrix is used to measure the performance of the CNN-LSTM based action recognition model. The prediction accuracy of each category is on the diagonal of the normalized confusion matrix. Note that a good classification model shall lead to a normalized confusion matrix, with diagonal elements as close to 1 as possible, and with off-diagonal elements as close to 0 as possible.

Confusion Matrix: A confusion matrix is a table that summarizes the performance of a classification model by showing the counts of true positive, true negative, false positive, and false negative predictions. It provides insights into the types of errors made by the model and helps in understanding its strengths and weaknesses.

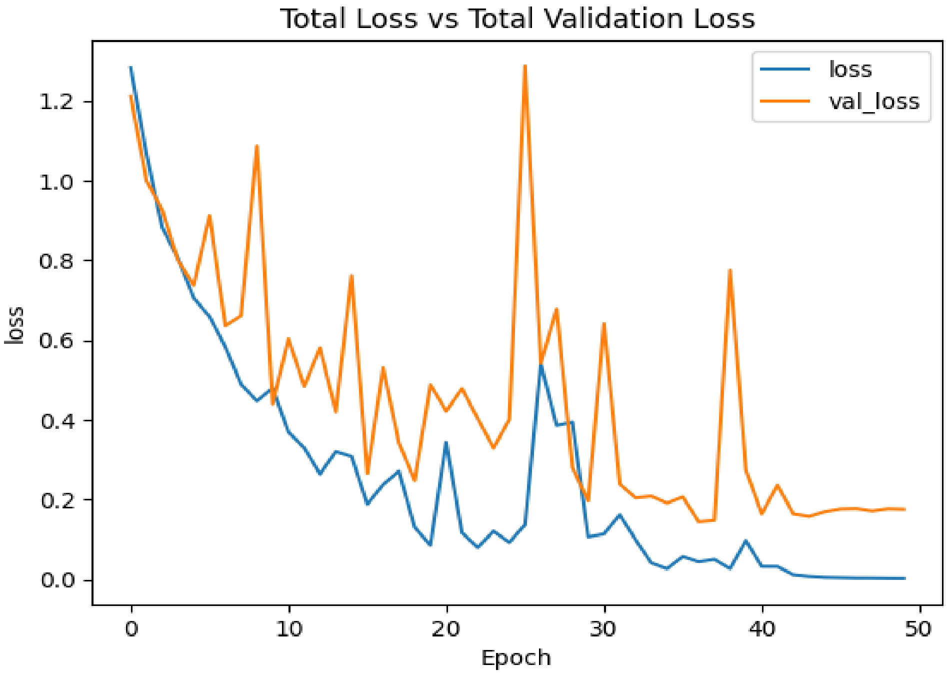


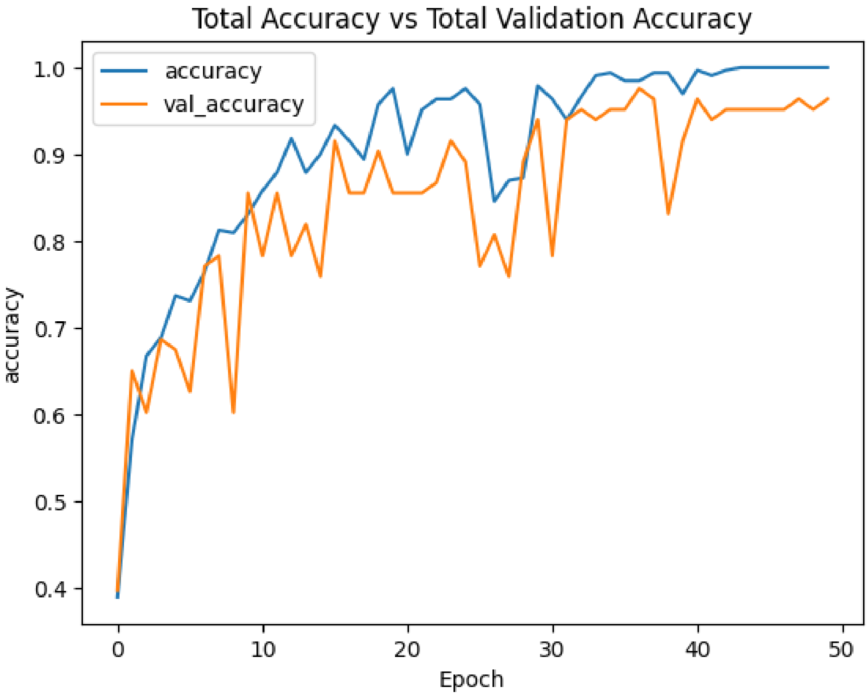
```
               precision    recall  f1-score   support

      Diving       0.86      1.00      0.92       516
    Drumming       1.00      1.00      1.00      1405
    HighJump       1.00      0.39      0.56       231
   HorseRace       1.00      1.00      1.00       504
 HorseRiding       0.99      1.00      0.99      2210
 JumpingJack       1.00      1.00      1.00       483
PlayingTabla       1.00      1.00      1.00      2943
   PoleVault       0.97      0.98      0.98      1520
 PommelHorse       0.96      1.00      0.98       275
     PushUps       1.00      1.00      1.00       216

    accuracy                           0.98     10303
   macro avg       0.98      0.94      0.94     10303
weighted avg       0.99      0.98      0.98     10303
```

RESULTS


Total Loss vs Total Validation Loss
loss
val_loss
loss
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
Epoch


Total Accuracy vs Total Validation Accuracy
accuracy
val_accuracy
accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
10
20
30
40
50
Epoch

CONCLUSION

Action recognition on video ,We have witnessed the dataset's significance, the power of CNNs and LSTMs, and the synergy of combining them. We encourage you to embark on your action recognition journey, experiment with CNN-LSTM models, and leverage the dataset to push the boundaries of AI.

Emotion Recognition: Understanding human emotions from facial expressions in videos requires analyzing both the spatial details of the face and the temporal evolution of expressions, making LSTM+CNNs valuable in this context.

These applications showcase the versatility of LSTM+CNN architectures in handling a wide range of tasks that involve spatiotemporal data. Their ability to combine spatial and temporal information makes them a powerful choice for tasks where context and timing are crucial.

REFERENCES

1. Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." Advances in neural information processing systems.

2. Donahue, Jeff, et al. "Long-term recurrent convolutional networks for visual recognition and description." Proceedings of the IEEE conference on computer vision and pattern recognition.

3. Wang, Limin, et al. "Temporal segment networks: Towards good practices for deep action recognition." European conference on computer vision. Springer, Cham.

4. Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.

5. Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." Proceedings of the IEEE international conference on computer vision.