A review of Deep Learning Models used for video analysis

Prof. Dishita Mashru, Dr. Komil Vora*, Dr Avani Vasant

Department of Information Technology,

V.V.P. Engineering College, Rajkot, Gujarat, India

Abstract:

The exponential growth in video data across domains like surveillance, sports, entertainment, and social media has intensified the need for efficient video analysis techniques. Deep learning has emerged as a transformative approach in automating video understanding, outperforming traditional computer vision methods. This paper presents a comprehensive review of various deep learning models—particularly CNNs, RNNs, hybrid architectures, and transformer-based models—used in video analysis tasks such as classification, human activity recognition, and object detection. The survey evaluates performance, model architectures, datasets, and key methodologies employed in recent research. We conclude by highlighting future research directions including multimodal fusion, real-time processing, and generalization to unconstrained environments.

Keywords: Video Analytics, Deep Learning, CNN, RNN, Video Classification, Human Activity Recognition, Transformer Models

Introduction:

In the digital era, videos constitute the majority of global internet traffic, driven by platforms like YouTube, TikTok, and surveillance systems. Traditional video analysis methods relied heavily on handcrafted features and rule-based logic, which often failed in dynamic, real-world scenarios. The advent of deep learning revolutionized this field by enabling automatic feature learning and end-to-end training pipelines.

Video analysis typically involves understanding both spatial (frame-wise) and temporal (time-based) features. While CNNs are powerful in learning spatial hierarchies from frames, RNNs (especially LSTMs and GRUs) and Transformers can model the temporal dependencies across sequences. Hybrid models combining these architectures have shown promising results in various domains such as sports analytics, surveillance, and healthcare.

This paper reviews state-of-the-art approaches employing deep learning models for video analysis. It focuses on model architectures, datasets, performance metrics, and comparative evaluation, offering insights for researchers and practitioners.

Literature Review:

Video analytics provides a broader area for research where there is an enormous amount of video data generated on the web with each passing day. In [1], the discussion about implementing the video analysis using CNN as well as RNN is mentioned. The utilization of COIN dataset that consists of 11,827 videos, which is divided into 180 different tasks are considered, and most of the videos are collected from YouTube. The key frames are extracted and selected based on several criteria like unicity of the frames, brightness score, entropy/ contrast score , histogram, etc. The proposed model took the input video, which was then preprocessed and passed for feature extractors and classifiers, for which the average of the predictions obtained was taken, and finally the relevant labels were applied to the predicted frames. This process is as shown in the figure below:

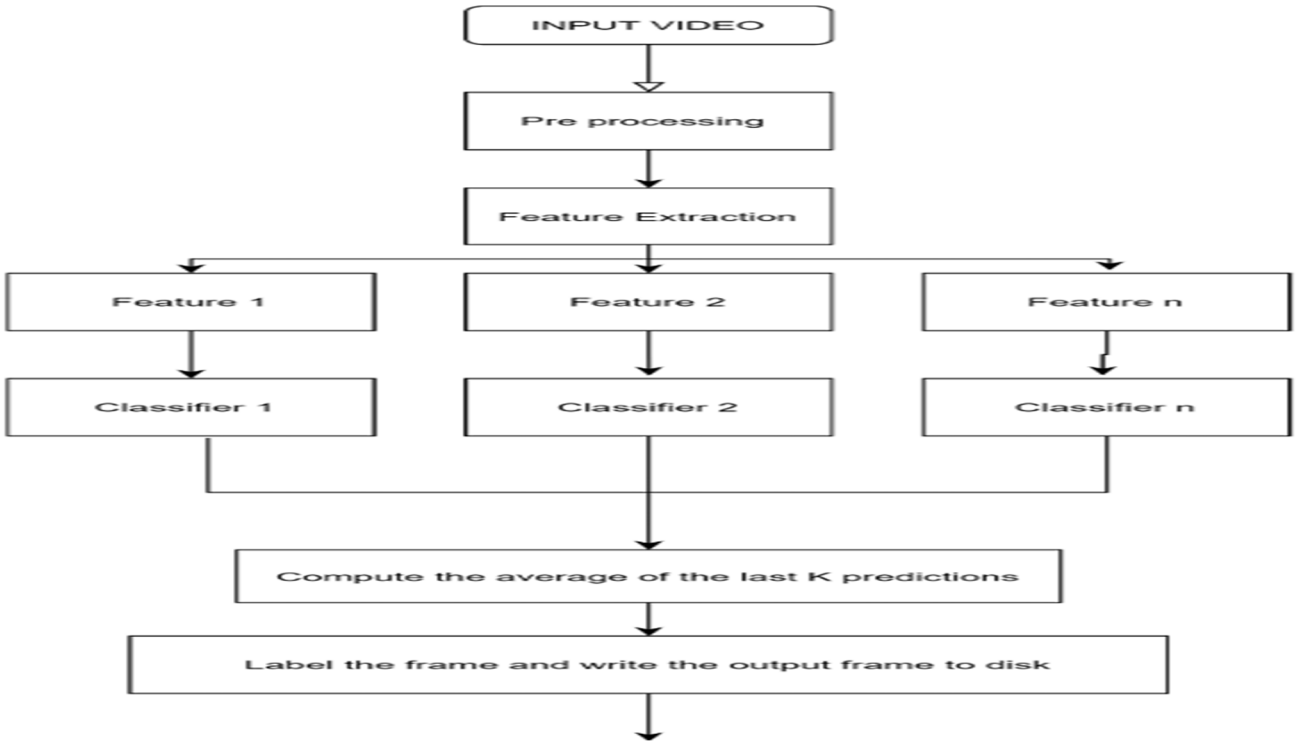

Fig. 1: Video Classification Model Used in [1]

The hybrid model of CNN as well as RNN was utilized in the video classification process. The accuracy of the model was found 80.27% , and the prediction results on the input video of cricket game was as shown in the below figure:

Fig 2: Prediction results obtained after deploying CNN+ RNN Hybrid Model [1]

A hybrid model consisting of the combination of CNN model and Video Transformer model (ViT) was proposed in [2] for Human Activity Recognition from videos. The proposed model is depicted in Fig. 3 as follows.

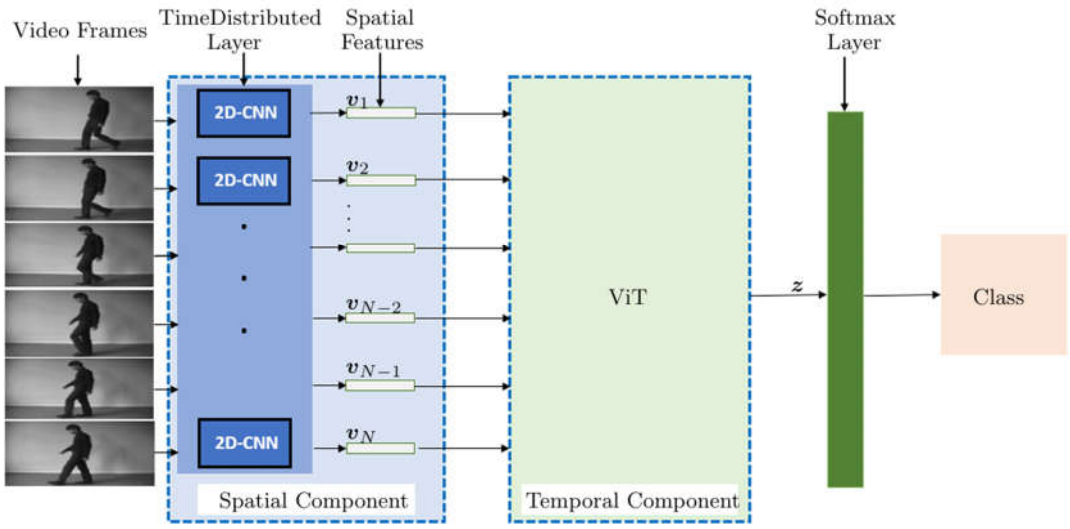

Fig 3: Proposed CNN+ViT model [2]

According to the proposed architecture, it included a Time-Distributed layer in addition to the CNN backbone, which is then followed by ViT model to identify the actions in the video input. The spatial component contains the MobileNet backbone of CNN in the Time Distributed layer to process each input frame individually, and provides the spatial features as output. The temporal component contains the ViT model to process the sequence of the spatial feature vectors obtained previously, processes it and then gives the final representation of the output, which is then provided as input to the Softmax layer to classify the action occurring in the input video. This model was trained using the KTH dataset, which comprised of six classes of actions, namely, walking, jogging, running, boxing, hand waving and hand

clapping, containing 2391 videos, processed at 25 frames per second. The accuracy of predicting the actions from the input video sequences was 97% for the given context.

The literature conducted from [3] described the survey on analyzing sports video by comparing the traditional neural network models with the deep learning model. It was observed that the deep learning models gave better accuracy as compared to the traditional models. The proposed model used in [3] is as depicted in the figure below:

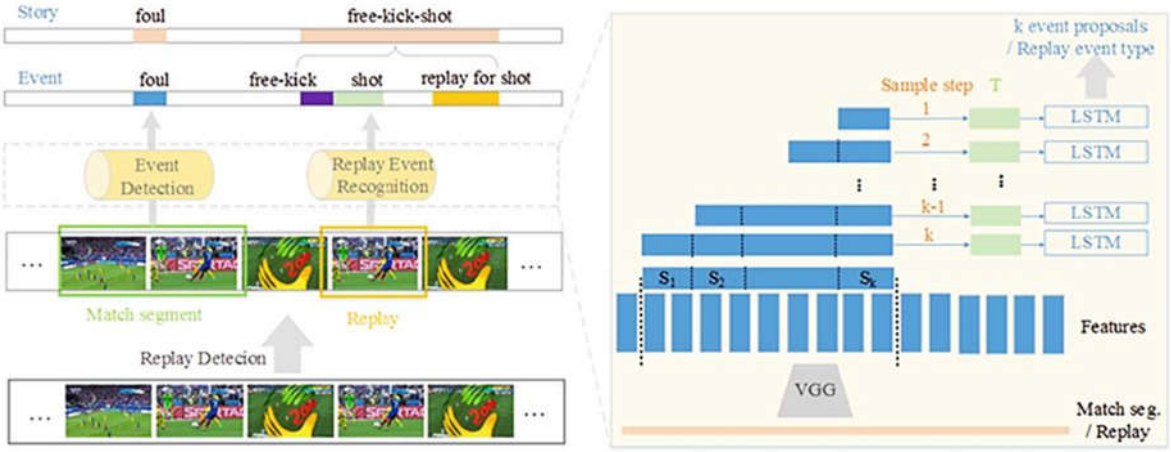

Fig. 4: Proposed work depicted in [3]

It was found that opting for the hybrid CNN and LSTM models for soccer activity recognition gave better prediction results and training as well as validation accuracies.

The survey described in [4] stated the advantages as well as disadvantages of the traditional video classification models, compared to the CNN and hybrid CNN- RNN models. The summary of the video classification techniques mentioned in its work is as depicted in the following figure. It also showed the study of various proposed models and provided a summarized table containing the models analyzed along with the respective accuracies, as shown in the table below:

Method	Accuracy
LRCN [48]	82.9
DT + MVSV [125]	83.5
LSTM–Composite [49]	84.3
FSTCN [126]	88.1
C3D [127]	85.2
iDT + HSV [128]	87.9
Two-Stream [61]	88.0
RNN-FV [129]	88.0
LSTM [50]	88.6
MultiSource CNN [130]	89.1
Image-Based [55]	89.6
TDD [35]	90.3
Multilayer and Multimodal Fusion [110]	91.6
Transformation CNN [131]	92.4
Multi-Stream [112]	92.6
Key Volume Mining [132]	92.7
Convolutional Two-Stream [62]	93.5
Temporal Segment Networks [39]	94.2

Table 1: Summary of the models studied in review carried out in [4]

The conclusion drawn from the literature review of [4] are: i) the hybrid architecture (CNN+RNN) performs better in the spatial feature extraction. ii) Models like LSTM and GRU can perform better than traditional RNN model, and better with hybrid CNN LSTM model.

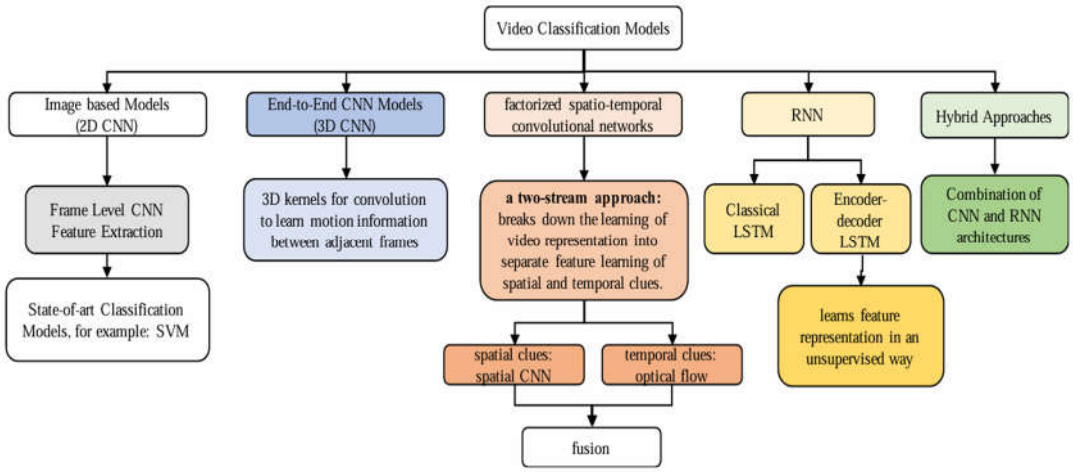

Fig. 5: Summary of video classification techniques mentioned in [4]

The proposed architecture in [5] uses the KTH dataset containing a large number of videos as mentioned above, from which the frames are extracted and preprocessing is done. The redundant frames are ignored and removed and then these frames are given as an input to the Pretrained InceptionV3 which is further passed to the LSTM model for training the dataset, that provides the classification results as an output. This is mentioned in the following figure. The accuracy of this model is reported to be 88.37%.

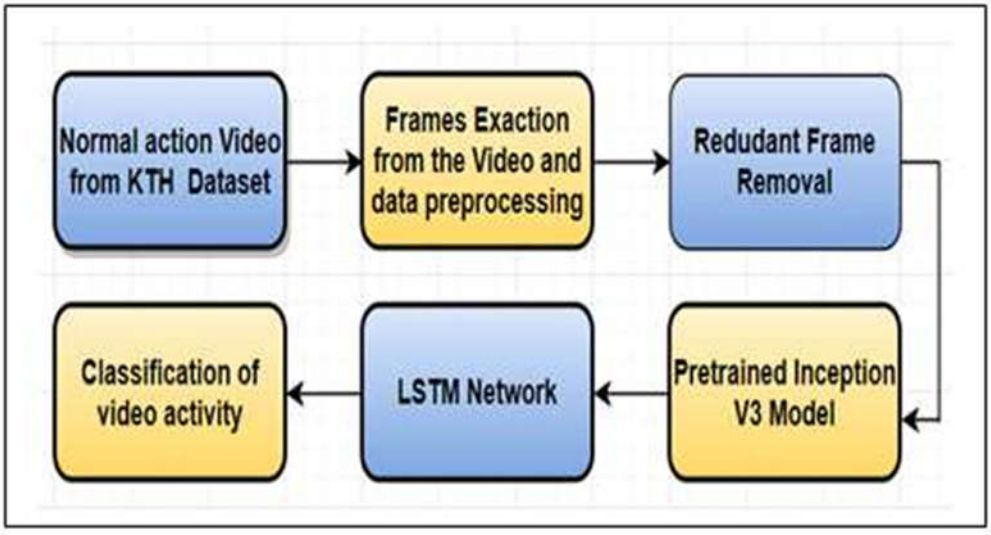

Fig. 6: Proposed InceptionV3+LSTM model in [5]

Comparative Evaluation:

Study	Model Architecture	Dataset	Accuracy	Key Feature
[1]	CNN + RNN	COIN	80.27%	Frame-level preprocessing
[2]	CNN + ViT	KTH	97%	Transformer-based sequence learning
[3]	CNN + LSTM	Soccer	-	Sports activity recognition
[4]	CNN, RNN, Hybrid	Various	-	Survey of models
[5]	InceptionV3 + LSTM	KTH	88.37%	Transfer learning

Table 1: Comparative Evaluation of Different Models

Conclusion:

This review paper was aimed to provide the basic idea of applying deep learning techniques for video analysis. The comparison between the traditional models defined in the papers referred here showed that the hybrid CNN and LSTM model gave better prediction and classification accuracies as compared to simple CNN or RNN or traditional models. For the future reference, we aim to enhance our research in the application of deep learning models in video analysis like Object Detection, Face recognition, gesture recognition, etc.

References:

[1] Patil, Pradyumn, et al. "Video content classification using deep learning." *arXiv preprint arXiv:2111.13813* (2021).

[2] Alomar, Khaled, et al. "RNNs, CNNs and Transformers in Human Action Recognition: A Survey and a Hybrid Model." *arXiv preprint arXiv:2407.06162* (2024).

[3] Rangasamy, Keerthana, et al. "Deep learning in sport video analysis: a review." *TELKOMNIKA* 18.4 (2020): 1926-1935.

[4] ur Rehman, Atiq, et al. "On the Use of Deep Learning for Video Classification." *Applied Sciences* 13.3 (2023): 2007.

[5] Begampure, Saylee, and Parul Jadhav. "Intelligent Video Analytics For Human Action Detection: A Deep Learning Approach With Transfer Learning." *IJCDS* 11 (2022): 63-72.

[6] Tran, Du, et al. "Learning spatiotemporal features with 3D convolutional networks." *ICCV* (2015).

[7] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? A new model and the kinetics dataset." *CVPR* (2017).

[8] Vaswani, Ashish, et al. "Attention is all you need." *NeurIPS* (2017).

[9] Feichtenhofer, Christoph, et al. "SlowFast networks for video recognition." *ICCV* (2019).

[10] Kay, Will, et al. "The Kinetics human action video dataset." *arXiv preprint arXiv:1705.06950* (2017).