
Proceedings of
INTERNATIONAL CONFERENCE
CHENNAI, INDIA
11th MARCH 2024
Organized by
ieeeforum
(Industrial Electronics and Electrical Engineers Forum)
Co-Organized by
Institute of
Research and Journals
Integrated Research and Innovation
In Association with
ISSRD
International Society for Scientific Research and Development
iFEARP World
irfsr
International Research Forum for Scientific Research

VIDEO CLASSIFICATION AND SIMILAR VIDEOS RECOMMENDATION SYSTEM

[1]VADDELLAMOHAN, [2]SEELASAIKUMAR, [3]THOTAHEMANTHKRISHNA, [4]KAKARLABHARGAV, [5]K.DEEPIKA

[1,2,3,4]Student,Dept.ofCSE(AI &ML),VasireddyVenkatadriInstituteofTechnology,AndhraPradesh,India.
[5]Assistant Professor, Dept. of CSE (AI &ML), VasireddyVenkatadri Institute of Technology, Andhra Pradesh,India.
E-mail: [1]mohanvaddella@gmail.com, [2]sai680513@gmail.com, [3]khemanth1107@gmail.com,
[4]kakarlabhargav220@gmail.com,[5]deepikak@vvit.net

Abstract - This work's main goal is to categorize videos, by extracting video frames and generating feature vectors to identify action sequences. For improved spatial and temporal properties, a suggested deep neural network integrates Long Short Term Memory and Convolutional Neural Network models. By removing characteristics from frames and taking into account variables like frame width, height, and video sequence duration, video categorization is accomplished. Beginning with a time-distributed 2D convolutional layer, Following the initial convolutional layer, time-distributed max-pooling layers and dropout layers are strategically inserted. Additional convolutional layers with increasing filter sizes further enhance the model's capacity for feature extraction. A flatten layer prepares the output for temporal analysis by a Long Short-Term Memory (LSTM) layer, which captures intricate temporal dependencies within the video sequences. The architecture culminates in a dense layer equipped with a softmax activation function, generating predictions for various classes. The ISRO Dataset is used for video classification. The results show that in terms of prediction accuracy, the LRCN methodology performs better than both ConvLSTM and conventional CNN methods. Furthermore, the proposed method offers improved temporal and spatial stream identification accuracy. The findings show that a 67% prediction of the action sequence's probability is made. Along with Classification, we also Recommend few videos based upon the input videos or images.

Keywords - Convolutional Neural Network (CNN), Long-term Recurrent Convolutional Network (LRCN), Long Short-Term Memory (LSTM).

I. INTRODUCTION

The rise of manufacturing in the digital age has reached unprecedented levels. Platforms like YouTube produce great videos and regular in-house content. They are often faced with the challenge of sorting through the amount of digital content. The research aims to address the challenge of developing classification models using Long-Term Recurrent Convolutional Networks (LRCN). It tries to better categorize videos, providing users with introductions and similar content. Our goal is to simplify the recovery process and provide a more efficient way to communicate large-scale digital video.

Our video classification system focuses only on content developed by the Indian Space Research Organization (ISRO). ISRO's video content is varied, including satellite launches, mission updates, and behind-the-scenes footage. Thus, a systematic approach to organizing and categorizing these videos is needed. Our research seeks to use LRCN and provide a solution to develop a robust model that can accurately classify ISRO-related videos. This targeted classification system is expected to enhance the accessibility and searchability of ISRO's extensive video archives and feed users seeking organization-related content the specific needs of the Organization.

By incorporating LRCN into our video classification model, we aim to achieve higher accuracy in discriminating between different content classifiers. This approach promises to help develop an effective monitoring system not only in ISRO but also as a comprehensive tool for video content processing in various industries. Through this research, we provide a seamless experience and easy-to-use in an ever-expanding industries. We want to use it so that video classification models can evolve.

Apart from Classification, we are also providing similar video for the user by using the Annoy Algorithm, which is like a recommendation system. This makes the user easier to look up at similar content as he/she was interested in. Many of the platforms like YouTube and Netflix are famous for their movie Recommendation Systems. Here with this Annoy Algorithm, we are providing the similar videos, we discuss it in deep in a while.

II. RELATEDWORK

1. Several studies have been conducted on video segmentation using visual features. Andrej Karpathy and others. [1] used a new dataset of 1 million YouTube videos in 487 categories to conduct a comprehensive empirical study of convolutional neural networks (CNNs) for large video classifications. The authors investigated different approaches enhanced the ability of CNN to capture local spatial and temporal information in time domain and proposed a multiresolution architecture as a possible way to speed up training. The performance of the proposed spatio-temporal network shows a

significant improvement from 55.3% to 63.9% compared to the robust feature-based baseline models but from 59.3% to 60.9% better performance compared to the single-frame model of the Improvement. Transfer learning was also applied to the UCF101 Action Recognition dataset.

2. The CNN (RCNN) iterative model was proposed by Ming Liang et al. [2] for object detection by adding recursive combinations to each convertible layer. Even though the input parameters are fixed, the behavior of RCNN clusters changes over time, so that the behavior of neighboring clusters influences the behavior of each cluster. This characteristic provides the ability of the model to provide information about context will be included, which is important for improving the discovery process. CIFAR-10, CIFAR-100, MNIST, and SVHN object recognition datasets were used to evaluate the model. On every dataset that is taken into consideration, it is found that RCNN performs better than cutting-edge models with fewer trainable parameters. These results demonstrate the benefits of conventional object recognition programs over sophisticated programs exclusively.

3. For gesture recognition in videos, Pigou, L. Et al. [3] presented a novel give-up-to-quit trainable neural network architecture that combines bidirectional recurrence and temporal convolutions with deep systems. The significance of repetition and its necessity, which includes temporal convolutions that result in significant profits, have been examined by the writers. We tested several approaches with the newly discovered outcomes from the Montalbano gesture reputation dataset.

4. Caleb Andrew et al. conducted surveys on the algorithms used in video classification techniques [4] to see whether the technique is more effective at predicting gestures or motions and can be applied to the categorization of action films. In deep learning models for visual sequence modeling, LSTM is crucial for the recognition of gestures and actions. Even though there are numerous movie categories, emerging methods frequently combine well-known categories to increase classification accuracy

5. A revolutionary neural network incorporating advanced convolutional layers and the remarkable long short-term memory (LSTM) was brought forth by Xia, K., et al. [5]. In order to gauge the true potential of this model, three publicly accessible datasets, namely UCI, WISDM, and OPPORTUNITY, were employed. Astonishingly, the model boasted an accuracy of 95.78 percent in the dataset from UCI-HAR, 95.85 percent in the dataset from WISDM, and a notable 92.63 percent in the OPPORTUNITY dataset.

III. DATASET

The dataset we have been working with is ISRO dataset. It contains videos divided into five different categories: Animation, Person Close Up, Graphics, Outdoor Launch Pad, and Indoor Control Room and the dataset contains 150 videos, with 30 videos in each category. The average length of the videos in the dataset is about 5 seconds.

Fig-1:ISRO videos: 5 videos for each category

The animation part consists of animated content, with visual effects and conceptual elements. The Person Close Up range focuses primarily on individuals in close-up shots. Graphics include videos that incorporate computer-generated images or graphics, which provide complex data or mental representations. The Outdoor Launch Pad section contains outdoor launches, spacecraft, rockets, or activities related to space exploration, providing insight into the fun side of space exploration. Finally, the Indoor Control Room section contains control room videos, monitoring stations, or similar furniture with Shan and the behind-the-scenes side of his scientific endeavors.

TotalActions	5Categories
TotalVideos	150
VideosperGroup	30
MinimumDuration	4sec
MaximumDuration	23sec
FramesRate	6Per second

Table1:Detailed view of Input

IV. PROPOSEDAPPROACH

The entire thing in this is, First the input video is converted into frames with a rate of 5 per second which is crucial for the Identifying of the continuous changes in the video. Now the each frame must follow the preprocessing stage, and the result of the preprocessing step is then forwarded to the LRCN Architecture and the architecture will be discussed in below. The result of the LRCN model is an vector representation of the input video then used for the

annoy algorithm which will identify by plotting the input video with the pre-trained models. So when we plot we can identify the nearest plotting are the Similar videos as the input video.

4.1 DatasetPreprocessing:

The initial step will be to import the ISRO video dataset, which will then be split into three sets: a training set, a test set, and a validation set, with ratios of 0.8, 0.1, and 0.1, respectively. During the pre-processing stage, a frame is first extracted from the video, then it is shrunk to measure 64 by 64, and finally, the pixel values are normalized. The training set will be put into the CNN and LRCN models after the dataset has been pre-processed. The model will then be tested using test data. Lastly, a comparison of accuracy will determine whether the model performs better.

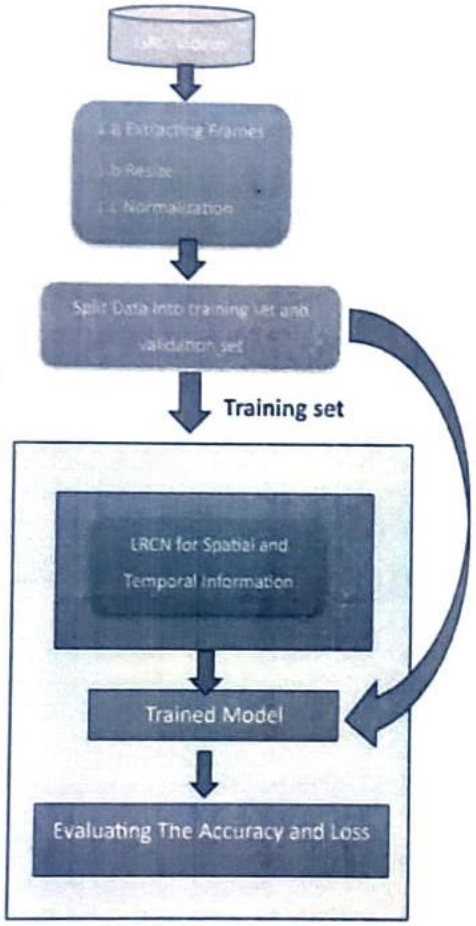

Fig-2:Illustration of Proposed Model

4.2 The Proposed Long-term Recurrent Convolutional Network For Video Classification

A sophisticated deep learning algorithm, Long-Term Recurrent Convolutional Networks (LRCN), can extract sequences of scene features from sequences of images. The versatility of the LRCN, which was first introduced [6], has been demonstrated in applications such as event recognition, image priming, and video annotation processing. This is possible because LRCN offers the flexibility of many recurrent neural networks (RNNs) are implemented. LRCN consists of two main components. Convolutional Neural Network (CNN), and RNN.

The main goal of a CNN encoder is to extract spatial features and transform them into one-dimensional vectors. The resulting vector is then used as input to the RNN encoder to make temporal dynamics decisions. Using shared loads allows the network to handle multiple frames simultaneously. The encoder-decoder design provides greater flexibility in the choice of architecture for CNNs and RNNs, since they operate independently and, because the returning neuron can receive input vectors of variable length, can be layered. The final output of any planar CNN architecture has served as input, regardless of its dimensions.

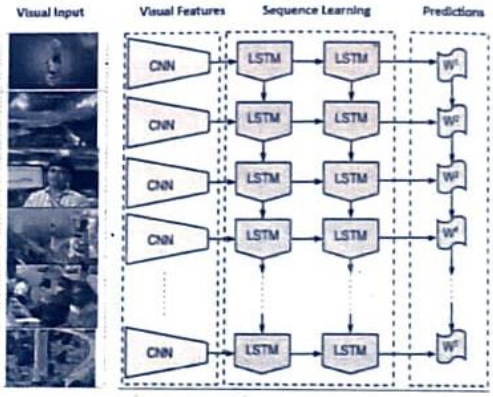

Fig-3: LRCN Architecture

4.3 Model size and Parameters of LRCN:

The Long-term Recurrent Convolutional Network (LRCN) combines Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), emphasizing Long Short-Term Memory (LSTM) networks. The version uses time-distributed CNN layers to process video frames. These CNN layers have 4 blocks, each with a Conv2D layer, BatchNormalization, MaxPooling2D, and Dropout layer. The Conv2D layers have filters starting from 32 to 128, the usage of a 3x3 kernel. MaxPooling2D reduces spatial dimensions, whilst Dropout layers counteract overfitting. After processing through CNNs, the output is flattened and exceeded to an LSTM layer with 64 units, ideal for coping with frame sequences through time. The version ends with a Dense layer using softmax activation for category, matching the number of trouble instructions.

V. ANNOY FOR SIMILAR VIDEOS RECOMMENDATION

A C++ library called ANNOY (Approximate Nearest Neighbour's Oh Yeah) has Python bindings that allow it to find points in space that are near a query point, such as a particular point of interest. Additionally, it builds sizable memory-mapped data structures that allow multiple processes to share the same set of data. Finding the points that are closest to a particular instance is the goal of the nearest neighbour search, a similarity issue.The nearest neighbours search determines the instance q that is closest to an instance p under a measurement function, given a set of n examples P = {p1, ..., pn} in some metric space X. Nearest neighbours search computes the closest instance q to an instance p under some measurement function.

The goal of the approximation closest neighbour search is to expedite the computation of the nearest neighbours, even while the exact method cannot be improved. Encouraging errors is the only way to speed up the calculation. Random projections and trees are the basis of the ANNOY (Approximate NearestNeighbour's Oh Yeah) algorithm. When using ANNOY, datasets with up to 100–1000 dense dimensions can be searched. The nearest neighbours are determined by dividing the collection of points in half. Until each set has k elements, this process is repeated. Generally speaking, k should be about 100 (see illustration below).

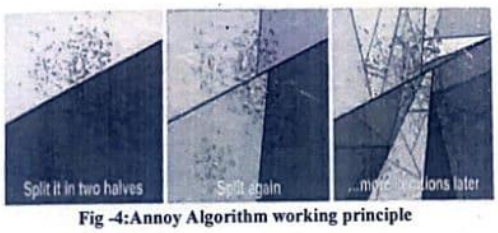

Fig -4:Annoy Algorithm working principle

Points whose distance from the query is at most c times the distance from the query to its nearest points can be returned using an approximate nearest neighbour search method. This method has the advantage that an estimated nearest neighbour is frequently nearly as good as an exact one. Specifically, minor variations in the distance should not be significant if the measure of distance effectively represents the concept of user quality. The output of an ANN's categorization indicates a class membership. Based on the majority vote of its neighbours, an object is classified. The object is classified into the class that its k closest neighbours share the greatest amount of. An integer k that is positive and usually tiny.

This is our ISRO dataset. The 150 films will also be plotted in the same way as the figure above, and each time a new input video data point is pointed, the annoy algorithm will produce 5 data points that fall under the same region according to the k value, for example, if k=5, then the annoy algorithm will provide 5 data points which fell under the same region.

VI. RESULTS

Throughout this research work, we work on one model is being evaluated on the previous mentioned ISRO dataset. We take care about the dataset avoiding the any sort of merging during the experiment process. The dataset was divided in a way that 80% was used means out of 150 means 120 videos had been used for the training of the model and the remaining 20% means 30 videos had been used for the testing the model. During the model initial phase, the validation split parameter was set to 0.2, representing that 20% of the training data was utilized for validationpurpose.

The results of our LRCN model described below Table Results represents that the LRCN model exhibits most impressive accuracy achieved is 67% on the new Video ISRO dataset. The "Training Loss" score reveals the degree of model error on the particular dataset it was trained on, whereas the "Validation Loss" score denotes the dégree of error on unseen, novel data. The fundamental purpose of training a machine learning model is to minimize its loss, which entails reducing the amount of error committed by the model.

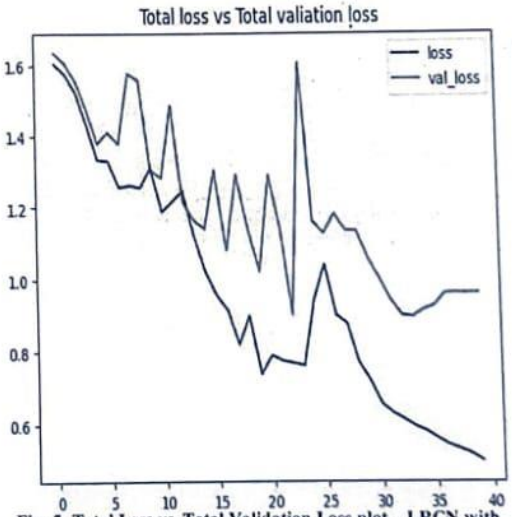

Fig- 5: Total Loss vs. Total Validation Loss plot – LRCN with the ISRO dataset.

"Training Accuracy" shows the model's performance on its training data, while "Validation Accuracy" indicates its capability on new data. A high validation accuracy is essential for effective model generalization, avoiding mere data memorization.

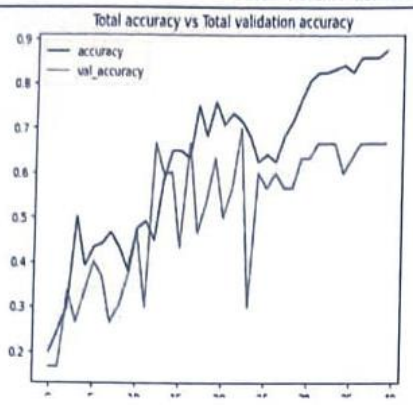

Fig-6: Total Accuracy vs. Total Accuracy Loss plot – LRCNwiththeISROdataset.

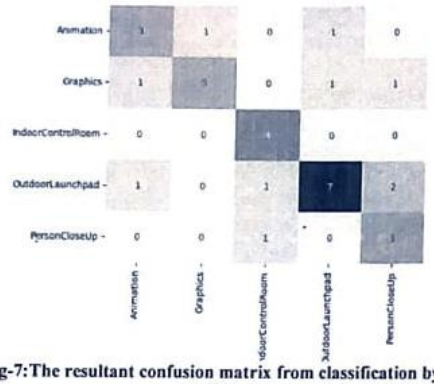

Fig-7:The resultant confusion matrix from classification by LRCN

A few misclassifications were observed from the above fig-7 such as a graphic video is identified as Outdoor Launch Pad and Outdoor Launch Pad is misclassified as Graphics. This misclassification reduces the accuracy of our LRCN model. The below fig-8 is an output where the given input video had been identified their belonging category and we can see the category label on the left corner with red text respectively.

Fig-8: Different Video Category Recognition Results.

VII. CONCLUSION

This study emphasizes the role of neural networks in video classification, underscoring the growing importance of such technologies. Utilizing Anaconda, TensorFlow, and Keras showcases the pivotal role of neural networks in this domain. It's suggested to integrate these technologies into educational programs for broader understanding. The LRCN model achieved a notable accuracy of 67%. And also seen how anoyalgorithm is well suitable for the videos recommendation. For future research, enhancing the ISRO dataset by including diverse ethnic backgrounds is essential for comprehensive representation. Adapting to such datasets will likely boost model accuracy, advancing activity recognition research. And here we had worked only on five categories so we can extend this model with the few other categories. Moreover, these models hold immense potential for practical applications. One such application could be their integration into Adult content detection and skipping that particular timeline. By analyzing the upcoming video frames from the video we can identify the adult content scenes based on the user interest he may skip the content by one click. In conclusion, by expanding and improving the dataset, refining the models, and exploring practical applications, the aim is to continue making significant contributions to the advancements in video category recognition technology.

REFERENCE

[1] Sanketh, S., Thomas, L., Rahul, S., Karpathy, A., George, T., & Li, F. (2014). Convolutional Neural Networks for Large-Scale Video Classification. In the IEEE Computer Vision and Pattern Recognition (CVPR) 2014 Conference Proceedings, Columbus, OH, USA (pp. 1725–1732).

[2] (2015) Ming, L. and Xiaolin, H. Convolutional Neural Network with Recurrent Architecture for Object Identification. Volume 3367–Volume 3375 of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, Boston, Massachusetts, USA.

[3] Dieleman, S., Van Herreweghe, M., Pigou, L., Van Den Oord, A., &Dambre, J. (2016). Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. 126(2-4), 430–439, International Journal of Computer Vision.

[4] Rex, F., and Caleb, A. (2018). A Survey on Action Recognition-Based Video Classification. 7(2.31), 89- 93, International Journal of Engineering & Technology.

[5] In 2020, Xia, Huang, and Wang published a book. LSTM-CNN Structure for Identifying Human Activity. Access IEEE, 8, 56855–56866.

[6] Mehr, H. D., &Polat, H. (2019). Human Activity Recognition in Smart Home With Deep Learning Approach. https://doi.org/10.1109/sgcf.2019.8782290

[7] Zhong-Qiu, Z., Xindong, W., Peng, Z., &Shou-Tao, X. (2019). A Review on Deep Learning for Object Detection. IEEE Transactions on Learning Systems and Neural Networks, 30(11), 1–21.

[8] Russo, M. A., Jo, K.-H., &Kurnianggoro, L. (2019). Classification of Sports Videos with Combination of Deep Learning Models and Transfer Learning. In Cox'Bazar, Bangladesh, proceedings of the 2019 International Conference on Electrical, Computer, and Communication

Engineering (ECCE), (pp. 1–5).

[9] Hood, S., Sayankar, B., and Baisware, A. (2019). Recent Developments in Video Data-Based Human Action Recognition are reviewed. The 9th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-19), held in Nagpur, India in 2019, has published its proceedings (1–5).

[10] Mengyun L. Deep Learning-Based Video Classification Technology. In Parallel and Distributed Systems (ISPDS), 2020 International Conference on Information Science, Xi'an, China, Proceedings (pp. 154-157).

[11] Deb, K., Dhar, P. K., Sarma, M. S., &Koshiba, T. (2021). Deep Learning Method for Classifying Traditional Sports Videos from Bangladesh. 11(5) Applied Sciences, 2149.

[12] In 2020, Roubleh, A. A., and Khalifa, O., O. Deep learning for video-based human activity recognition. The 7th International Conference on Electronic Devices, Systems, and Applications (ICEDSA2020) was held in Shah Alam, Malaysia, and the proceedings are available on pages 020023-1 through 020023-8

★★★