

Classification of Videos using Neural Network Technique

Anusha K M¹, Bhumika D M², Nithya H S³, Hemanthkumar S S⁴ and Dr. Sharath M N⁵

¹⁻⁴UG Student, Computer Science and Engineering, Rajeev Institute of Technology, Hassan, India
Email: anushakm.9590@gmail.com, dmbhumika1712@gmail.com, nithyhs23@gmail.com, hemanthkumars398@gmail.com

⁵Assistant Professor, Computer Science and Engineering, Rajeev Institute of Technology, Hassan, India
Email: sharathmn64@rithassan.ac.in

Abstract—Exercises involving video classification are now more frequently used. The topic has attracted more attention, especially since deep learning models proved to be a useful tool for automatically classifying videos. In order to acknowledge the significance of video classification tasks and summarise the efficiency of deep learning models for this task, this paper provides an incredibly thorough and concise review of the field. There are currently a number of reviews and survey papers on video classification in the scientific literature. The most recent state-of-the-art works, however, are not included in the current review papers because they either have limitations or are out of date. This paper provides an updated and succinct review by highlighting the key findings based on the current deep learning models. To identify potential areas.

Index Terms— Automatic video classification ,Video Processing.

I. INTRODUCTION

The feature that distinguishes video-to-text recognition as a subdomain of vision-related applications is the capacity to identify and recognise the actions or goals of the agent, which can be any object or entity that performs an action with end goals. Agents can be a single agent carrying out the action, a group of agents carrying out the action, or agents interacting in some way. One example of an agent in this sense is the human being itself, which is aware of human activity. The goal of human video-to-text recognition is to automatically analyse ongoing activities from an unknown video or a series of image frames. The system's objective in a simple scenario where a video is segmented to only include one instance of a human activity is to precisely place.

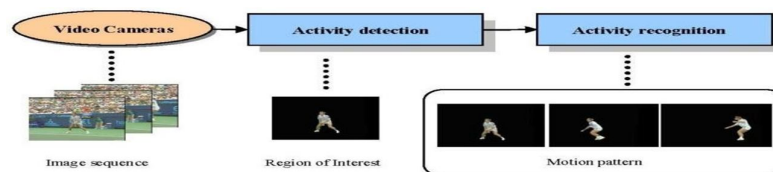


Figure 1: Flowchart for Hand Gesture Recognition

II. LITERATURE SURVEY

A brand-new DBN-based HAR technique that aims to improve the accuracy of human activity categorization was proposed by Abdellaoui, M., Douik, et al. [1]. As a first step, divide the video clips from the human activity dataset into frames. After that, the result is converted to binary frames, which are then enhanced in quality using a variety of morphological filtering methods. Then, after converting the new frames into a binary vector, create an input matrix containing the training data, testing data, and their labels.

A more effective technique for locating temporal activity in video was put forth by Yu-Wei Chao et al. [2] and is known as TAL-Net. It was based on the Faster RCNN object identification framework. Through the use of a multi-scale architecture that can handle extreme

Additionally, the deep learning models have particularly demonstrated a notable improvement in the classification of videos with an increase in the availability of massive datasets. W. Samke and others [5] This paper reviews studies that use deep learning techniques to classify videos. The scientific literature currently contains a number of reviews and survey papers that deal with video classification. However, those review papers either have some restrictions or are outdated and do not include the most recent state-of-the-art works.

The techniques covered in this review are not well developed, but Q. Ren [6] recently conducted a brief review on video classification methods. described, and the review also falls short of outlining research gaps, benchmark datasets, restrictions on current techniques, and performance.

III. CHALLENGES

1. Identifying Concurrent Activities: Concurrent activities are those that one or more people are performing at the same time. One user might interact with another user while sipping coffee, for instance. It is difficult to process videos to recognise both activities.
2. Recognising Interleaved Activities: People conduct activities in unpredictable ways in real life. People may decide to alternate between two or more different activities. Interleaved activities are what are known as such activities. As an illustration, the actor might take a call or watch TV while preparing food.
3. Ambiguity in interpretation: The method for identifying the activity should be adaptable to account for all these factors since different actors have different ways of carrying out a given action. A system should be able to recognise all different walking styles as "walking" activities without hesitation because different people have different walking habits.
4. Multiple resident complexity: Recognising all of the actions taken by each actor when multiple actors are present in the same scene becomes challenging. There may be circumstances where certain actors are only dimly or imperfectly visible.
5. With so many potential applications, there is room for precise and straightforward systems.

IV. PROPOSED METHODOLOGY

A. Convolution Layer

From an input image, it is the first layer to extract features. Applying filters to the convolution of an image allows for operations like edge detection, blurring, and sharpening. A convolutional layer in a neural network can be created by combining a number of convolutional filters. The convolution layer preserves pixel relationships by extracting image features from tiny input data squares. An image matrix and filters or a kernel, for example, are two inputs that are used in this mathematical operation.

B. ReLU

It is a linear rectified unit. It is customary to use a nonlinear layer (also known as an activation layer) right away following each convolution layer. During the convolution layers, the system had essentially only been computing linear operations; the goal of this layer is to introduce nonlinearity to the system. ReLU layers perform significantly better because the network can train much more quickly (due to the computational efficiency) without significantly affecting accuracy.

C. Pooling layer

Programmers might prefer to add a pooling layer after a few ReLU layers. It is helpful to reduce the size of the activation map in the deep end of the network to make the network easier to manage for classification. In general, the deep layers of the network require more filter matrices to recognise multiple high-level patterns but less information about the precise spatial locations of features.

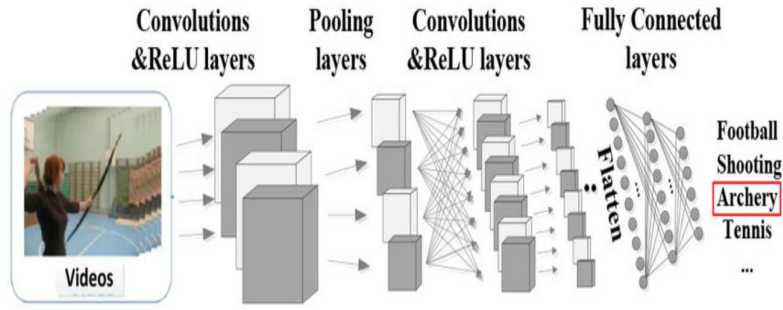


Figure 2: 3D-CNN architecture to classify videos

D. Fully connected Layer

A set of fully connected layers make up an absolutely connected neural network. Every input dimension affects every output dimension. The fully connected layer determines which feature needs to be connected to a specific task by analysing the output of the previous layer. A connected layer determines which feature needs to be related to a specific task by analysing the output of the layer above it. A fully connected layer basically determines which high-layer feature strongly correlates with a specific class and has specific weights so that when you compute products between the weights and consequently the previous layer, you get the correct possibilities for the various categories.

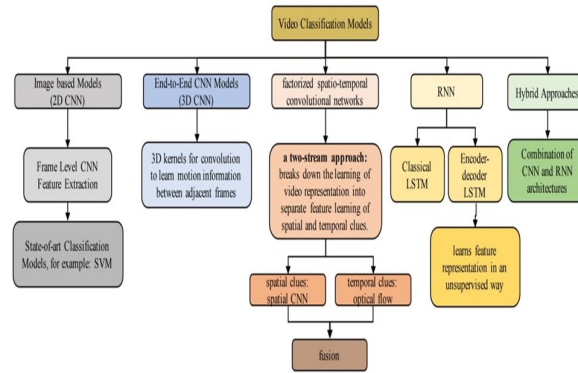


Figure 3: An overview of video classification techniques

V. ENCODING ARCHITECTURE

$$\mathbf{h}_t = \mathbf{f}(\mathbf{W}^{(hh)}\mathbf{h}_{t-1} + \mathbf{W}^{(hx)}\mathbf{x}_t) \quad (3.5)$$

Encoder Architecture

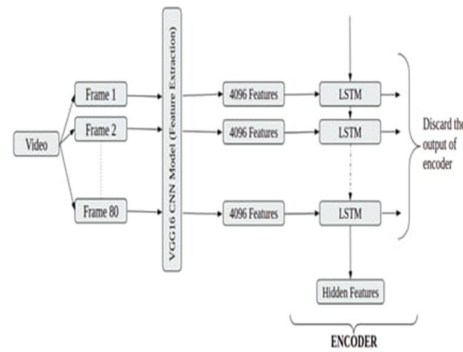


Figure 4: Encoding Architecture

This is the model's final hidden state as produced by the encoder. Equation 3.5 can be used to ascertain it. This vector makes an effort to incorporate all of the data from the input elements in order to aid the decoder in producing accurate predictions. It serves as the model's initial hidden state for the decoder.

VI. DECODING ARCHITECTURE

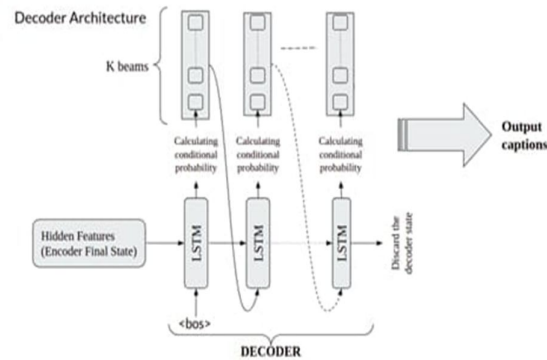


Figure 5: Decoding Architecture

A collection of recurrent units that each predict an output (y) at a particular time step t . Each recurrent unit takes a hidden state from the unit before it and outputs it along with its own hidden state. The output sequence includes the captions that were created for the frames.

VII. RESULTS AND DISCUSSION

Here is an example of the program's actual outputs as determined by the video and image inputs, as displayed below.



Figure 6: Feature Extraction

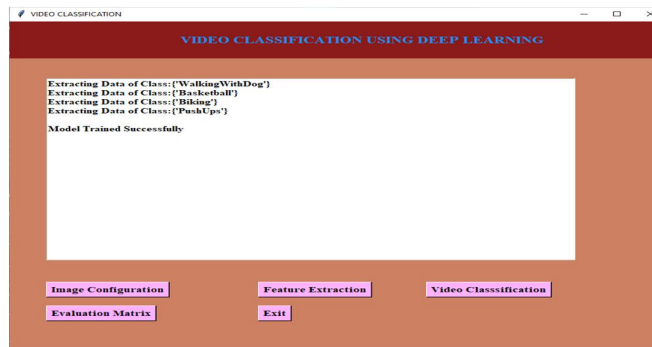


Figure 7: Evolution Matrix

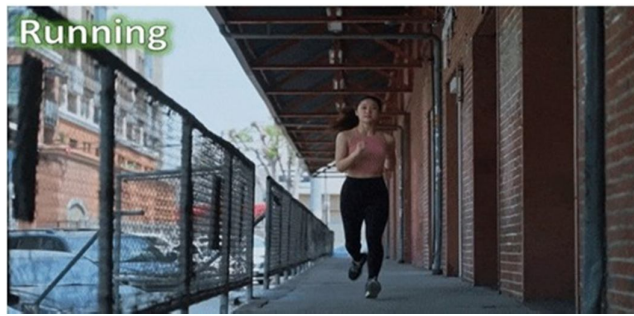


Figure 8: Output Image

VIII. CONCLUSION AND FUTURE SCOPE

In this project, we can classify the video and identify human activity based on the actors' behaviour. Predicting the various tasks that people carry out each day is our aim. The CNN algorithm and machine learning algorithms were used to implement this model. Using neurons, which are the most effective at prediction and classification, we have increased the model's efficiency. The main objective is to offer a solution for optimisation.

In order to provide a comprehensive solution that is independent of datasets and appropriate for the real world, we first suggested concentrating on open-set action recognition in future work. Second, transfer learning can be used in conjunction with already-known architecture information to improve feature extraction. Third, more hybrid approaches need to be presented in order to optimise and improve the deep learning paradigm and fully utilise various types of architecture.



Figure 9: Future work's potential research directions

REFERENCES

- [1] "Fuzzy logic-based human activity recognition in video surveillance applications," *Adv. Intell. Syst. Comput.*, vol. 427, pp. 227-235, 2016. doi: 10.1007/978-3-319-29504-6_23. S. Abdelhedi, A. Wali, and A. M. Alimi.
- [2] Faster RCNN object identification framework, Yu-Wei Chao et al., "TAL-Net," vol. 43, no. 3, 2011. doi: 10.1145/1922649.1922653.
- [3] "Spatiotemporal multiplier networks for video action recognition," in *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-Janua, pp. 7445-7454, by C. Feichtenhofer, A. Pinz, and R. P. Wildes.
- [4] "A review on video classification with methods, findings, performance, challenges, limitations, and future work," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 6, no. 2, p. 47, 2021.
- [5] "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247-278, 2021; Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K. R. Müller.
- [6] Q. Ren et al., "A Survey on Video Classification Methods Based on Deep Learning," *DEStech Trans. Computer Science and Engineering*, no. cisnrc, 2019.
- [7] A. Anushya, "Video Tagging Using Deep Learning: A Survey," *Int. J. Computer Science and Mobile Computing*, vol. 9, no. 2, pp. 49-55, 2020.
- [8] P. Rani, J. Kaur, and S. Kaswan, "Automatic Video Classification: A Review," *EAI Endorsed Trans. Creat. Technol.*, vol. 7, no. 24, p. 163996, 2020.
- [9] Y. Li, C. Wang, and J. Liu, "A systematic review of literature on user behaviour in video game live streaming," *Int. J. Environ. Res. Public Health*, vol. 17, no. 9, 2020.