International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Hybrid Deep Learning Techniques for Large-Scale Video Classification

Saif Saad Alnuaimi[1*], Bilal Hikmat Rasheed[2], D. Yuvaraj[3], P. Sundaravadivel[4], R. Augustian Isaac[5]

Submitted: 10/12/2023 **Revised:** 22/01/2024 **Accepted:** 01/02/2024

Abstract: Effective large-scale video management and classification are becoming more and more necessary due to the Internet's video data rapidly increase. A comprehensive evaluation of the trade-off between timeliness and efficacy should be made during real-world implementation. In industrial deployments, the frame extraction function is frequently used to categorize video actions, while the video classification technique integrated with a time segment network is implemented. The scientific literature now contains several reviews and research articles on the topic of video categorization. With the ability to analyze spatial and temporal information concurrently and efficiently, the combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) provides an effective framework for video categorization issues. This research proposes a comparison to evaluate how CNNs and RNNs integrated into different architectures might use temporal information to enhance video classification accuracy using deep learning. To optimize the performance of the proposed design for a CNN and RNN hybrid that works well, an innovative action template-based feature extraction technique is presented. This approach extracts features by analyzing the similarity between each frame's informative areas. Using RNN based video classifiers extensive experiments were performed on the UCF-50 and UCF-101 datasets. The efficiency of the suggested Feature extraction technique is demonstrated by the considerable improvement in video categorization accuracy shown in the experimental data, as examined by a one-way statistical evaluation of variance.

Keywords: *Deep Learning, Video Classification, Convolutional Neural Networks, Recurrent Neural Networks, Feature Extraction*

1. Introduction

The process of automatically determining the classification of an input video is known as video classification. Along with serializing and comprehending the spatiotemporal connection within the image, the main issues include predicting classifications using the extracted model and abstracting the most significant data [1]. Due to its extensive use in numerous domains, including video surveillance, video analysis, and video retrieval, and video annotation, video classification has become a popular study area in computer vision. Significant advancements in video classification using various learning models have been achieved in recent years. Nevertheless, the video experience overflows make it challenging to outperform earlier approaches using a single view feature [2]. Moreover, different inter-class changes and fine-grained intra-class variations might lead to misclassification and confusion in classification.

Due to advancements in network designs, storage capacities, and accessibility to digital cameras particularly those found in mobile phones, video has grown in prevalence in various fields in recent years [3]. Over 500 hours of video have been uploaded to the Internet every minute, according to current information. The demand for video content is predicted to drive an enormous rise in the amount of videos over the next decades. Because of this the development is noteworthy and presents significant difficulties for systems that organize, store, and retrieve videos. On social networking platforms, films mostly focus on human behavior. To ensure the effective use and administration of these movies, automatic semantic content classification is crucial. The complexity of video data makes it difficult to classify video content, too.

Deep learning techniques have been used to solve action recognition issues in the image and video domains. In the last ten years, convolutional neural networks, or CNNs, have generated innovative outcomes. CNN applications are not as effective in the video domain as they are in the image domain. Recurrent neural networks (RNNs) have thus been used to improve video classification efficiency by gathering temporal information through sequence learning. Considering the advantageous results of integrating CNNs and RNNs [4, 5], the representation of temporal information remains a difficult issue because of the intricate changes that occur in behaviors and the dynamic context of videos. The integration of additional training data and transfer learning has resulted in a significant improvement in action recognition ability. Many video datasets, including UCF-101 [6], have been made available, and subsequent reports on these benchmarking datasets have demonstrated the most recent findings [7–11].

[1]*Department of Computer Science,*
Cihan University-Duhok, Duhok, Iraq.
Email: saif.saad@duhokcihan.edu.krd
**(Corresponding Author)*
[2]*Department of Computer Science,*
Cihan University-Duhok, Duhok, Iraq.
Email: bilal.rasheed@duhokcihan.edu.krd
[3]*Department of Computer Science,*
Cihan University-Duhok, Duhok, Iraq.
Email: d.yuvaraj@duhokcihan.edu.krd
[4]*Department of Artificial Intelligence & Machine Learning, Saveetha Engineering College, Saveetha Nagar, Thandalam, Chennai, India.*
Email: sundar.me2009@gmail.com
[5]*Department of Artificial Intelligence & Machine Learning, Saveetha Engineering College, Saveetha Nagar, Thandalam, Chennai, India.*
Email: sangam.naadu@gmail.com

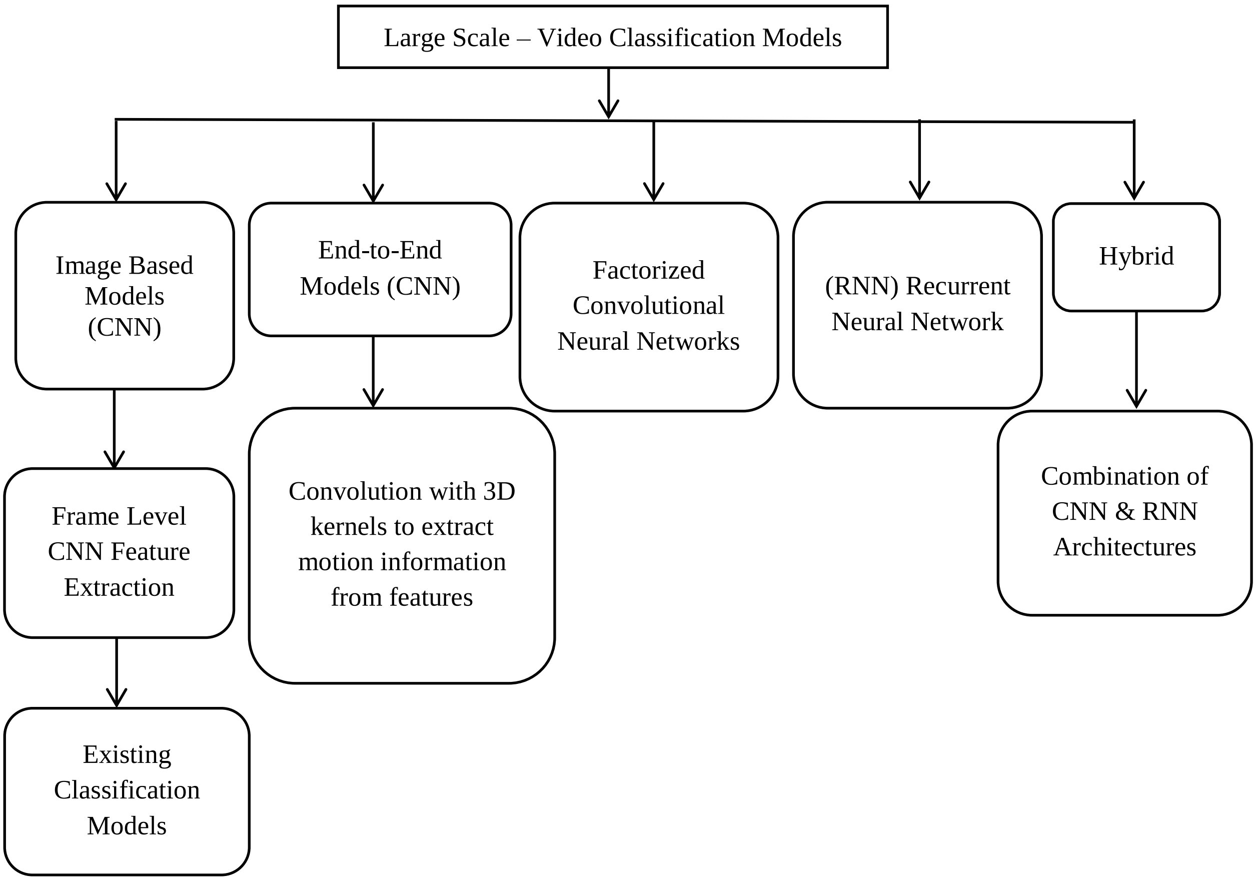

Large Scale – Video Classification Models
Image Based Models (CNN)
End-to-End Models (CNN)
Factorized Convolutional Neural Networks
(RNN) Recurrent Neural Network
Hybrid
Frame Level CNN Feature Extraction
Convolution with 3D kernels to extract motion information from features
Combination of CNN & RNN Architectures
Existing Classification Models

Fig. 1.1. A Summary of Methods for Video Classification

Most existing video categorization techniques categorize videos by giving each frame a label. However, since some frames contain more distinguishing information than others, evaluating all frames identically degrades the classification performance. They suggest that to improve classification performance, Feature selection is crucial. To maintain the basic content, this research suggests an innovative Feature extraction technique that involves defining a procedure template that represents the whole film in a series of Features. The suggested Feature extraction technique, which uses a scenario template for each video to identify and choose the most unique frames with both static and dynamic frames without requiring for complicated procedure is the primary innovation of this study. A Summary of Methods for Video Classification is shown in Figure 1.1.

The two primary achievements of this work are the formulation of the action template-based Feature extraction proposal, which is designed to get more relevant frames by determining similarities only between action zones, rather than whole frames, and the finding of the optimum design for combining CNNs and RNNs for video classification. The prior work, which has been much developed, was partially published in [12] and is used as a benchmark to evaluate the recently suggested approach. Numerous tests conducted for this work have demonstrated an important superiority of the action template-based Feature extraction approach over the frame selection techniques tested in our experiments for comparisons.

This is how the remainder of the paper is structured: Section II reviews prior work, while Section III describes the suggested hybrid deep learning-based extraction approach. The performed experiments and the results are examined and summarised in Section IV, and Section V contains the conclusions.

2. Related Works

Researchers presented several highly developed and deep-net-based action recognition techniques during the last ten years. Handcrafted features for non-realistic action videos have been aided by earlier works. Since the deep neural network (DNN) is the basis of the suggested approach, the authors will solely review relevant works that promote DNN in this part. Various deep learning algorithm variations have been presented recently for computer vision challenges such as human activity recognition in videos, and have demonstrated excellent results.

Techniques for extracting Features might be broadly divided into six categories: motion analysis-based, visual content-based, shot boundary-based, uniform sampling-based, and clustering-based. Uniform sampling-based approaches, while simple and computationally efficient, might not accurately describe the video in two different situations: either there are insufficient Features for a brief, semantically significant movie, or there are too many Features with comparable material for a long, static portion [13]. Shot boundary-based approaches were the main focus of early Feature extraction efforts. In fundamental terms, this method uses shot boundary detection to identify the beginning or middle frame of every shot as the Feature. In [14] examined approaches for video shot boundary detection. Shot boundary-based techniques are simple to use and broadly applicable, however, the retrieved Feature is unstable and unable to fully capture the visual content.

Features are chosen using a shot activity-based technique, which takes advantage of the frame that differs from the others the least from a certain similarity metric. Using the premise that "every Feature corresponds to a contiguous interval in a shot," In [15] developed a Feature selection technique founded on this idea. The boundaries of intervals and the Feature's position within each interval are optimized in this work. Similar to this, [16] uses the Lloyd–Max technique to create a scalar quantize. Feature-based video summarization and information retrieval based on visual content have both been investigated using this visual content-based technique. Features in movie segments are analyzed using this method, which extracts visual information from video clips. Utilizing video content data derived from a parsing procedure, this study presented an integrated system solution. Using Feature extraction, a semantic video summary was also created by simulating the human attention process.

A description called attention quantifier, which identifies color prominence and mobility with greater attention concerned, is used to quantify the visual attention of each frame [17]. Numerous attempts have been made to examine the visual aspects of videos to identify Features to divide and summarize videos [18]. Using optical flow computations to identify local minima of action in a single shot, a unique algorithm was suggested for choosing Features inside shots from video using a motion analysis-based method [19]. Optical flow analysis is used in this work to measure motion in a shot, with Features chosen at the local minima of the activity. A method for action Feature identification that involves the L1-norm and accumulated optical fluxes has also been developed. A similar method that uses optical flow and mutual information entropy calculation was previously found for salient region-based Feature extraction [31].

Feature extraction has been accomplished using clustering-based techniques. The aim is to use a clustering algorithm such as K-means to group frames depending on their low-level attributes, and then select the most identical images with the centers of the groups [20]. Features have also been extracted using dynamic Delaunay graph clustering using an iterative edge reduction method. By using K-means and the artificial fish swarm methods to extract Feature sequences, Tan et al. showed the KGAF-means technique [21]. This research proposes a strategy for dealing with several significant drawbacks of the previously discussed methodologies. While shot-based methods for Feature extraction are user-friendly, they are not sufficient to preserve the temporal information. Regarding clustering-based methods, their temporal complexity is substantial and their sensitivity depends on the kind of kernel used and the number of clusters.

Moreover, video is a unique sort of media content with intricate backdrop information and temporal information. Managing complete frame variations as compared to a particular region of interest is a further disadvantage of the previously described techniques. To overcome these constraints, this research suggests an innovative technique that depends on the similarity between regions of interest in multiple frames. In contrast to other research, our study employs an action template to identify each video's region of interest. Significantly, our prior research [22] provided some work related to deep neural networks for video classification.

3. Proposed Work

One of the main pre-processing steps in video analysis is feature extraction. The goal of feature extraction is to efficiently obtain more discriminating information from the video. Every video has distinct qualities of its own, including brightness, contrast, saturation, blur, vibration, camera angle, number of performers, action type, length, and background. A significant flaw in feature extraction results from considering all movies similarly and taking into account a lot of characteristics in each one. Therefore, in a continuous action film, it is imperative to identify the activity region. This work provides a new feature extraction method based on the difficult task of locating the action, given the fluctuations in the complicated video data.

3.1. Feature Extraction

In most cases, a reviewer's center of interest on the screen and camera corresponds to the action's position in an image. When analyzing and documenting, it has been noted that the central region receives the majority of attention. Consequently, video frames are typically divided off once the region of interest is identified. Next, a framework for the video to follow the action area is created by defining the area of activity as a region in the middle of the frames that results in either the most variation or the least similarity among later frames. Feature extraction is made better and more precise by minimizing the impact of potentially unpredictable backgrounds by evaluating only the variation in regions of activity between frames across the images [3].

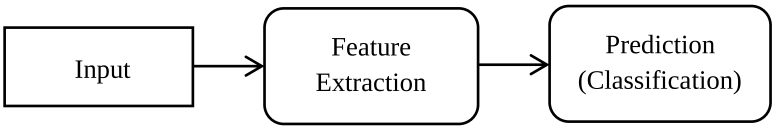


Fig. 3.1. Procedure for Classifying Videos

The four steps in the suggested feature extraction process are as follows: A pre-defined amount of features in chronological sequence are chosen, (1) an action template is identified, (2) the action's position is specified, (3) action similarities are calculated to locate distinctive frames, and (4) an action template is found. Figure 3.1 shows the process of classifying videos.

3.2. Video Classification

The DL algorithms used in the video classification challenge are reviewed in comprehensive detail and effectively in this section. This section includes an overview of video data techniques, classic manual methods, innovations in video classification, and the most recent state-of-the-art DL algorithms for video classification.

3.2.1. Convolutional Neural Networks (CNNs)

Matrix multiplication is used in Convolutional Neural Networks (CNNs) to produce outputs that can be used in subsequent training cycles. Convolution is the term for this technique. Convolutional neural networks are the name given to this kind of neural network because of this. Word vectors are the representation of words in a sentence or news article in natural language processing (NLP). A CNN is then trained using these word vectors. By giving a kernel size and several filters, the training is done. There are multiple dimensions in a CNN [23].

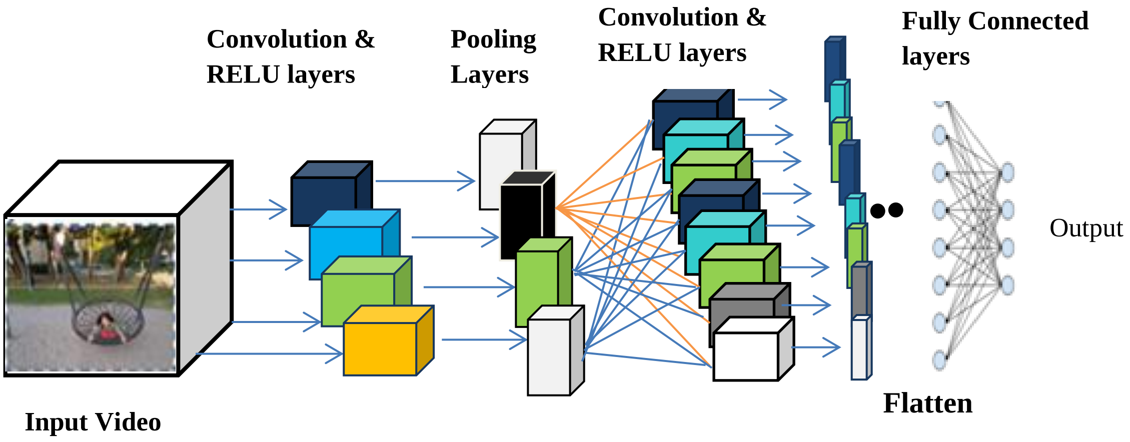


Fig. 3.2. An Illustration of CNN Architecture to Classify Videos

One-dimensional CNNs are typically utilized in text classification or natural language processing applications. CNN works with word vectors represented by one-dimensional arrays. A CNN uses a fixed-size window filter to run over the training set. At every phase, the filter weights multiply the input, producing an output that is saved in an output array. The data's feature map or output filter is represented by this output array. This method uses the input training data to identify a feature. Kernel size determines the filter's size, and the number of filters indicates how many feature maps will be employed. CNN may be used to discover local features that are obtained straight from the training set in this format. Figure 3.2 shows the representation of CNN architecture of video classification.

3.2.2. Recurrent Neural Networks (RNNs)

Data is processed sequentially in a Recurrent Neural Network (RNN) to facilitate learning. The capability of this sequential system to retrieve the sequence that occurred before the one it is processing now validates it. Because the output at each time step is used as the input for the following time step, it is known as recurrent. The preceding time step's output is recalled to accomplish this. Thus, our system can discover long-term dependencies in the training set.

In the framework of NLP, learning about several news components in connection with one another is preferable to learning about each article independently. Memory cells are arranged in layers that make up an RNN. RNNs can use a variety of memory cell types. The Long Short-Term Memory (LSTM) unit or cell is one example of this type [24]. When the sequence is processed at each time state, LSTM comprises a carry, a cell state, and the current word vector in process. Providing that there is no loss of information along the sequential procedure is the responsibility of the carrier. An LSTM cell is made up of three distinct learning gates and weights. An input gate processes the current input, an output gate predicts values, and a forget gate eliminates extraneous data at each time step.

3.2.3. Hybrid CNN and RNN Techniques

The proposed framework combines the LSTM's capacity to learn long-term dependencies and the CNN's capacity for extracting local features. Initially, the input vectors are processed using a CNN layer of Conv1D to identify the local features that are found at the text level. The CNN layer's output serves as the input for the LSTM units'/cells' subsequent RNN layer. Using the local features that were collected by CNN recovered, the RNN layer analyses the long-term dependencies of the local features in news items to determine whether they are phony or real.

As CNN-RNN can capture both local and sequential properties of input data, it has demonstrated effectiveness for several classification and regression problems. Utilizing their capacity to acquire image characteristics by the CNN and sequence features with the RNN, they have been utilized, for instance, in emotion identification [25] and in recognizing signs from video streams [26]. According to [27] and [28], RNNs may handle spatial relations well enough to discover long-term connections between text entities and significant attributes. It may also extract temporal and context features from text in NLP applications.

Deep learning algorithms are beneficial, but they also have certain practical drawbacks. These include the need for large training datasets, the challenge of determining the ideal hyper-parameters for every issue and dataset, and the lack of interpretability. These drawbacks directly affect the models' efficacy in new and unrecognized operations, leading to their behavior as "black-box" oracles [29]. The foundation of next-generation DL optimizing techniques is established by recent developments in bio-inspired techniques, which enable the enhancement of DL characteristics. The suggested hybrid approach aims to advantage a great deal from hyper-parameter optimization, and it is a component of the next research in this area that aims to investigate the different bio-inspired methods and determine which is most suited for the given task [30].

4. Experimental Results and Evaluation

To confirm the model's validity, an experimental evaluation was performed during the following section. The data set and experiment conditions are described in detail first, followed by an analysis of the model's output. Finally, a comparison is made between the technique's efficiency and the state-of-the-art approach to video classification.

4.1 Dataset Description

The UCF50 [31] dataset is made up of 6,618 real-world films that were labeled with 50 action types, ranging from everyday activities to general sports, and were obtained from YouTube at a fixed frame rate of 25 frames per second. These videos are divided into 25 groups, and videos within a category may have comparable content—for example, the same person, background, or point of view.

Datasets for video categorization that are widely used include UCF101 [32]. 13,320 video clips that are divided into 101 classifications constitute this UCF50 addon. Performing musical objects, sports, body-motion only, human-human communication, and human-object communication are five classifications into which these 101 classifications may be divided. These video segments run for more than 27 hours in total length. Each video has a resolution of 320 × 240 and is gathered from YouTube at a constant frame rate of 25 frames per second. The standard setting of three train-test splits [33], is followed for the division of training and test sets.

4.2 Performance Metrics

Four metrics, which depend on the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) in the binary classifiers' predictions, were implemented to evaluate the results:

- The proportion of True (i.e., accurate) forecasts is called accuracy.
- Recall measures the classifier's capacity to identify every positive sample.
- The precision of the classifier refers to its capacity to identify a negative sample as positive.
- Scores in the dimension [0, 1] are computed for the $F1$ score, corresponding to the harmonic mean of precision and recall.

The metrics are computed by the following formulae:

Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$

Precision = $\frac{TP}{TP+FP}$

Recall = $\frac{TP}{TP+FN}$

F1- Score = $\frac{2*(\text{Precision} \times \text{Recall})}{Precision+Recall}$

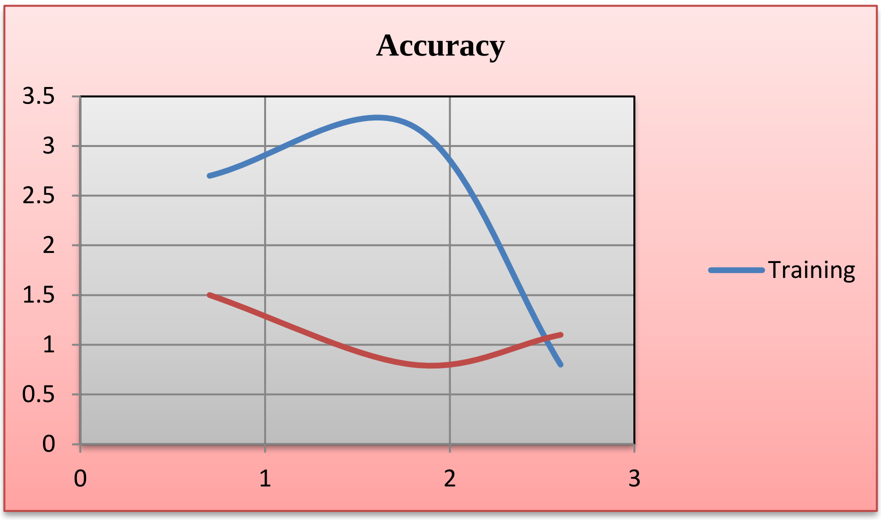
After each iteration of the evaluations and providing accuracy at 98% confidence intervals, a paired t-test was applied to verify the statistical importance of the findings.

The hold-out validation method is used for parameter adjustment during the training phase. The validation scores are used to determine the optimal parameters. Next, by forecasting the classes of test images that haven't been viewed yet, the algorithm with the most effective features is determined using testing data. Implementing the CUDA v9.0 toolset and TensorFlow-GPU v1.12 on an NVIDIA Titan X GPU, the suggested network designs are executed into execution. The neural network optimizer is used to minimize costs, and the batch size is set to 128. By early stopping, the number of epochs is calculated by tracking the variation in validation loss. As a regularisation technique, dropout disables certain neurons with a probability of 0.5 in CNN and RNN.

F1-Scores are generated for analyzing performance in the evaluations, and accuracy is applied to compare the performances developed by various architectures. The UCF-101 organization has provided three training-testing splits, with 180 training, validation, and testing accuracy ratings gathered. Furthermore, using the standard training, validation, and test splits of the UCF-50 dataset, 90 accuracy ratings were obtained. The evaluation of normality can be made numerically by comparing the calculated cumulative distribution with the cumulative distribution that's likely to arise if the data had a normal distribution. This comparison is done using the Kolmogorov-Smirnov test, a normality test. Regarding the test of homogeneity of variances, the dependent variable's variances are homogenous given the mean and median, as demonstrated by the application of the Levene statistics.

Levene's test, which is suitable for testing the null hypothesis, is essentially a single-way comparison of variance on the standard deviation of the variations between every result and the average of its category. To evaluate variance ratios and determine whether the findings are important, an ANOVA test has also been performed. The Tukey's honest significant difference (HSD) test has then been performed to ascertain whether the means of the respective groups are distinct. In general, the hybrid CNN-RNN approach that has been suggested outperforms all other supervised classification techniques.

Further tests were conducted utilizing the hybrid CNN-RNN approach, which was trained on the UCF-100 dataset and tested on the UCF-50 dataset via an identical configuration, to advance the work in this field. Given that numerous algorithms outperform the 0.9 classification accuracy threshold, UDF-100 is selected for training due to its larger size and less potential for enhancement. Plotting the training and validation accuracy and loss values throughout 10 epochs, Figure 4.1 illustrates the UCF-100 trained algorithm's capacity for generalization on a different dataset. The findings indicate that despite the training accuracy and loss finding their peak after six epochs, the validation accuracy stays relatively constant across all epochs and is not as high as what was attained during training on the UCF-50 dataset.



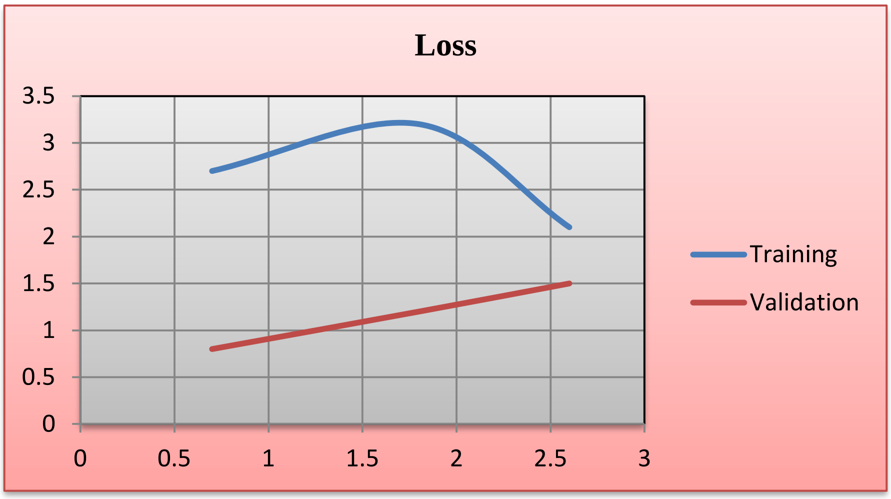
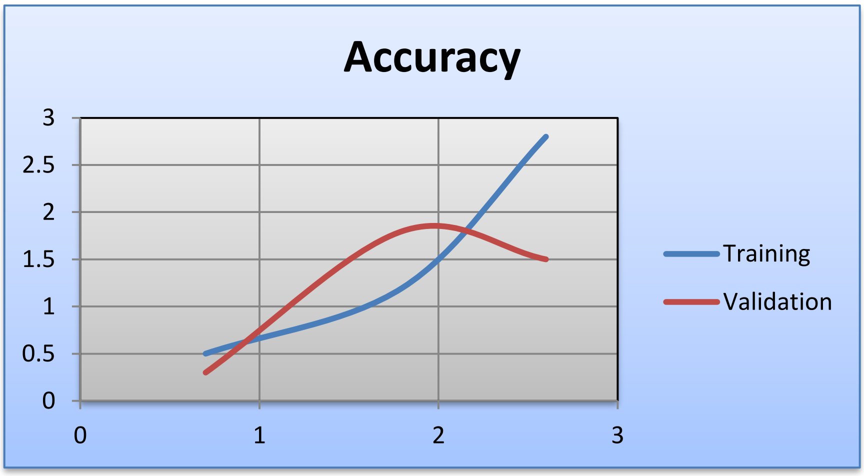


Fig 4.1. Accuracy and loss representations for validation of standardization training.

Given that the loss varies significantly with an increasing pattern, the validation loss plot indicates that the algorithm may have been over-fitted. The algorithm outperforms poorly on the fake news dataset with the same structure, considering the comparatively large corpus used for training and the model's near-1.0 accuracy on training, indicating poor generalization in the cross-dataset validation results [34]. The capacity of the framework to generalize could likely be enhanced by the addition of an internal drop-out layer, and this should be further investigated in the upcoming research on the subject. The learning curves of the suggested approach on the two datasets, as well as the training and validation loss and accuracy curves across the 10 epochs on each dataset, are provided to help determine whether the algorithms are over-fitted to the data. For this work, the PyPlot feature of the Matplotlib toolkit was applied, and the outcomes are shown in Figures 4.2 and 4.3.



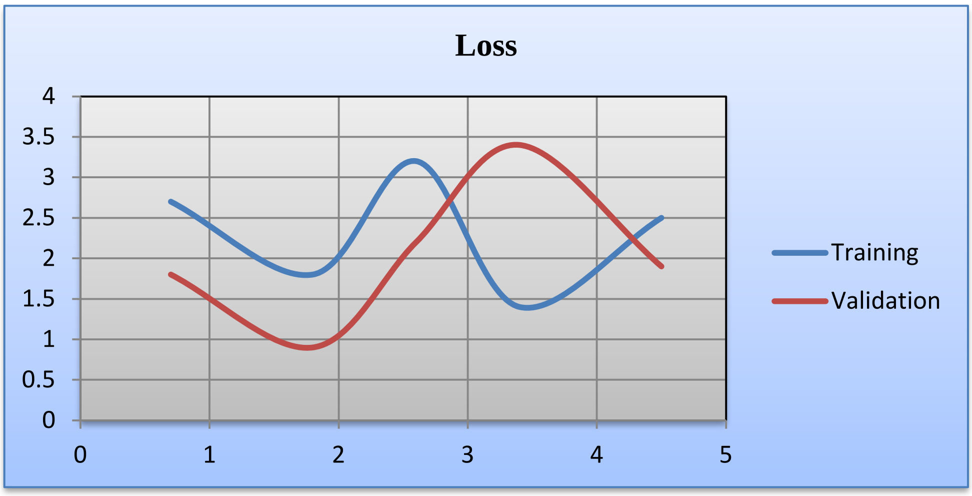
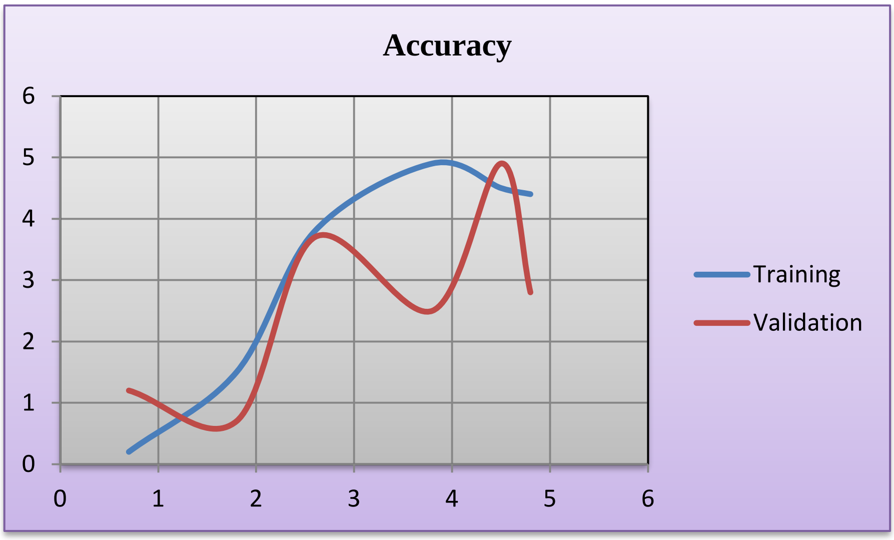


Fig. 4.2. UCF-50 Accuracy and Loss Charts for Training and Validation.

The proposed approach demonstrated high accuracy on the ISOT dataset. The framework improves in its ability to categorize the content, as shown by the training and testing accuracy increasing with the epochs and correspondingly decreasing loss rates. The training and validation accuracy for the FA-KES dataset does not grow smoothly. Likewise, the validation loss is approximately the same during the 10 epochs. Over-fitting is a real-world consequence of training a model without considering its capacity for generalization. The poor predictive accuracy on a new, unexplored dataset is the most evident effect of over-fitting. Furthermore, over-fitted algorithms reveal high levels of complexity and analyze far more data than is likely required for establishing an evaluation. Lastly, over-fitted systems reduce reusability because they need to be retrained from scratch and cannot be transferred to a comparable task on a different dataset.

Deep learning also has a functional consequence in that it makes people less proficient at other jobs because of their concentration in certain areas. The suggested hybrid approach overcomes the limitation of neural networks to processing one challenge at a time by combining an RNN which archives the text's sequential flow with a CNN network and analyses the text's spatial, or conceptual, properties. The fundamental hypothesis is empirically demonstrated hybrid Convolutional Neural Networks (CNN) Recurrent Neural Networks (RNN) architecture might enhance existing baselines for video classification. The hybrid standard techniques yield much worse performance than the experiments conducted on two distinct real-world UCF-50 and UCF-100 datasets. At last, the enormous amount of training data may aid deep learning systems. Several human-annotated datasets that may be utilized to train an algorithm for successful video classification have been examined in this work. If many datasets are fed into and formatted correctly, a generalized algorithm that may recognize the characteristics of each dataset and determine what is legitimate and what is not can be trained on them.



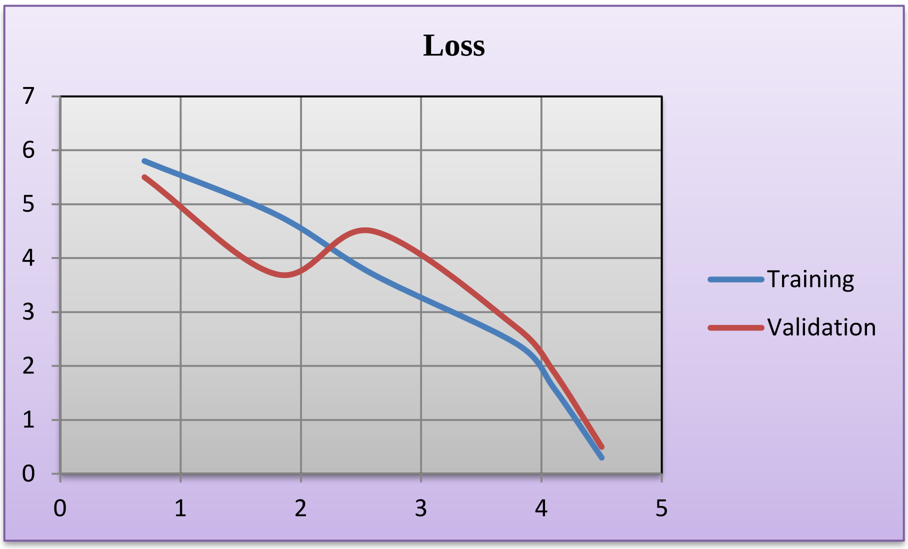


Fig. 4.3. UCF-50 Accuracy and Loss Charts for Training and Validation

5. Conclusion

This work proposes a template-based feature extraction technique that uses action template-based similarities to extract features for tasks related to video categorization. The combination of pre-trained CNN has outperformed the other architectures in terms of classification accuracy. It is demonstrated that dynamic background noise is efficiently excluded from being regarded as actions in feature selection by computing structural similarities between two important regions of successive frames. The analysis and experimental findings demonstrate that the suggested feature extraction strategy may accurately identify informative frames, hence improving the deep learning frameworks' performance for video categorization. At last, the suggested approach effectively extracts appropriate features from human action films for deep neural network-based video categorization by identifying the pertinent region with the retrieved action template. This study has some drawbacks, even though the suggested approach performed better than the two popular feature extraction techniques. One of the drawbacks is that the CNN framework that was implemented to evaluate the suggested approach was not the most recent model. This indicates that the outcomes were not optimal.

The experiment's inadequate technological setup, consisting of a single GPU machine, prevented the study from conducting more extensive tests with higher groups, which is an additional deficiency. On two distinct datasets, the suggested feature extraction technique executed more effectively than the widely used feature extraction techniques. Subsequent research endeavors may concentrate on implementing the suggested technique through more robust structures for practical video classification and summarising issues.

References

[1] Gong, X., & Li, Z. (2022). A Video Classification Method Based on Spatiotemporal Detail Attention and Feature Fusion. *Mobile Information Systems*, *2022*.

[2] Hu, Z. P., Zhang, R. X., Qiu, Y., Zhao, M. Y., & Sun, Z. (2021). 3D convolutional networks with multi-layer-pooling selection fusion for video classification. *Multimedia Tools and Applications*, *80*, 33179-33192.

[3] Savran Kızıltepe, R., Gan, J. Q., & Escobar, J. J. (2023). A novel keyframe extraction method for video classification using deep neural networks. *Neural Computing and Applications*, *35*(34), 24513-24524.

[4] Ballas, N., Yao, L., Pal, C., & Courville, A. (2015). Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*.

[5] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).

[6] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

[7] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299-6308).

[8] Duan, H., Zhao, Y., Xiong, Y., Liu, W., & Lin, D. (2020, August). Omni-sourced webly-supervised learning for video recognition. In *European Conference on Computer Vision* (pp. 670-688). Cham: Springer International Publishing.

[9] Kalfaoglu, M. E., Kalkan, S., & Alatan, A. A. (2020). Late temporal modeling in 3d cnn architectures with bert for action recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16* (pp. 731-747). Springer International Publishing.

[10] Mao, F., Wu, X., Xue, H., & Zhang, R. (2018). Hierarchical video frame sequence representation with deep convolutional graph network. In *Proceedings of the European conference on computer vision (ECCV) workshops* (pp. 0-0).

[11] Qiu, Z., Yao, T., Ngo, C. W., Tian, X., & Mei, T. (2019). Learning spatio-temporal representation with

local and global diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12056-12065).

[12] Savran Kızıltepe, R., Gan, J. Q., & Escobar, J. J. (2019). Combining very deep convolutional neural networks and recurrent neural networks for video classification. In *Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part II 15* (pp. 811-822). Springer International Publishing.

[13] Lin, J., Gan, C., & Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7083-7093).

[14] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6202-6211).

[15] Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., & Gong, B. (2021). Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16020-16030).

[16] Wang, X., Xiong, X., Neumann, M., Piergiovanni, A. J., Ryoo, M. S., Angelova, A., ... & Hua, W. (2020). Attentionnas: Spatiotemporal attention cell search for video classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16* (pp. 449-465). Springer International Publishing.

[17] Ahmad, H., Khan, H. U., Ali, S., Rahman, S. I. U., Wahid, F., & Khattak, H. (2022). Effective video summarization approach based on visual attention.

[18] Apostolidis, E., Balaouras, G., Mezaris, V., & Patras, I. (2021, November). Combining global and local attention with positional encoding for video summarization. In *2021 IEEE international symposium on multimedia (ISM)* (pp. 226-234). IEEE.

[19] Wu, G., Lin, J., & Silva, C. T. (2022). Intentvizor: Towards generic query guided interactive video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10503-10512).

[20] Ghauri, J. A., Hakimov, S., & Ewerth, R. (2021, July). Supervised video summarization via multiple feature sets with parallel attention. In *2021 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6s). IEEE.

[21] Bao, G., Li, D., & Mei, Y. (2020, September). Features extraction based on optical-flow and mutual information entropy. In *Journal of Physics: Conference Series* (Vol. 1646, No. 1, p. 012112). IOP Publishing.

[22] Nguyen-Thai, B., Le, V., Morgan, C., Badawi, N., Tran, T., & Venkatesh, S. (2021). A spatio-temporal attention-based model for infant movement assessment from videos. *IEEE journal of biomedical and health informatics*, *25*(10), 3911-3920.

[23] Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), 100007.

[24] Graves, A., & Graves, A. (2012). Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37-45.

[25] Kollias, D., & Zafeiriou, S. (2020). Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing*, *12*(3), 595-606.

[26] Masood, S., Srivastava, A., Thuwal, H. C., & Ahmad, M. (2018). Real-time sign language gesture (word) recognition from video sequences using CNN and RNN. In *Intelligent Engineering Informatics: Proceedings of the 6th International Conference on FICTA* (pp. 623-632). Springer Singapore.

[27] Zhang, X., Chen, F., & Huang, R. (2018). A combination of RNN and CNN for attention-based relation classification. *Procedia computer science*, *131*, 911-917.

[28] Zhou, C., Sun, C., Liu, Z., & Lau, F. (2015). A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630*.

[29] Drumond, T. F., Viéville, T., & Alexandre, F. (2019). Bio-inspired analysis of deep learning on not-so-big data using data-prototypes. *Frontiers in computational neuroscience*, *12*, 100.

[30] Kar, A. K. (2016). Bio inspired computing–a review of algorithms and scope of applications. *Expert Systems with Applications*, *59*, 20-32.

[31] Reddy, K. K., & Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine vision and applications*, *24*(5), 971-981.

[32] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

[33] Wu, Z., Wang, X., Jiang, Y. G., Ye, H., & Xue, X. (2015, October). Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 461-470).

[34] Muruganandam, S., Joshi, R., Suresh, P., Balakrishna, N., Kishore, K. H., & Manikanthan, S. V. (2023). A deep learning based feed forward artificial neural network to predict the K-barriers for intrusion detection using a wireless sensor network. *Measurement: Sensors*, *25*, 100613.