IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 3 January 2024, accepted 24 January 2024, date of publication 1 February 2024, date of current version 8 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3360929

RESEARCH ARTICLE

Integration of Feature and Decision Fusion With Deep Learning Architectures for Video Classification

RUKIYE SAVRAN KIZILTEPE[1], JOHN Q. GAN[2], AND JUAN JOSÉ ESCOBAR[3]

[1]Department of Software Engineering, Karadeniz Technical University, 61080 Trabzon, Turkey
[2]School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, U.K.
[3]Department of Software Engineering, CITIC, University of Granada, 18071 Granada, Spain

Corresponding author: Rukiye Savran Kiziltepe (rukiye.savrankiziltepe@ktu.edu.tr)

This work was supported in part by the Ministry of National Education, Turkey.

ABSTRACT Information fusion is frequently employed to integrate diverse inputs, including sensory data, features, or decisions, in order to leverage the advantageous relationships among various features and classifiers. This paper presents a novel approach for video classification using deep learning architectures, including ConvLSTM and vision transformer based fusion architectures, which incorporates the combination of spatial and temporal features, along with the utilisation of decision fusion at multiple levels. The proposed vision transformer based method uses a 3D CNN to extract spatio-temporal information and different attention mechanisms to pay attention to essential features for action recognition and thus learns spatio-temporal dependencies effectively. The effectiveness of the methods proposed in this paper is validated through empirical evaluations conducted on two well-known video classification datasets, namely UCF-101 and KTH. The experimental findings indicate that the utilisation of both spatial and temporal features is essential, with the superior performance gained by using temporal features as the primary source of features in conjunction with two types of distinct spatial features when compared to other configurations. The multi-level decision fusion approach proposed in this study produces results comparable to those of feature fusion methods while offering the advantage of reduced memory requirements and computational costs. The fusion of RGB, HOG, and optical flow representations has demonstrated the best performance compared to other fusion methods examined in this study. It has also been demonstrated that the vision transformer based approaches significantly outperformed the ConvLSTM based approaches. Furthermore, an ablation study was conducted to compare the performances of vision transformer based feature fusion approaches for enhancing the performance of video classification.

INDEX TERMS Computer vision, data fusion, deep neural networks, human action recognition, spatio-temporal features.

I. INTRODUCTION

Recently video classification has been remarkably advanced through deep learning, with a variety of applications including action recognition, gesture recognition, anomaly detection, and surveillance. Accurately classifying video data is of great importance for these applications, which is the foundation of automated systems that interpret complex dynamic visual information [1]. Video classification presents several challenges, including various lighting and viewpoint in video scenes, and the existence of motion information in large-volume video data.

Researchers have proposed various approaches for video classification to address these problems. Extracting spatial and temporal features from videos is a popular method. Temporal features capture the motion and dynamics of objects across frames, whereas spatial features capture the appearance of objects and scenes in each frame.

Many studies have attempted to gather the semantics of videos by leveraging Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), due to the emergence and availability of deep learning models

The associate editor coordinating the review of this manuscript and approving it for publication was Janmenjoy Nayak.

and large labelled video datasets, in which one of the main concerns is the integration of appearance and motion data within deep learning architectures. For this purpose, information fusion techniques have been commonly applied to effectively combine multiple data sources in video classification architectures, which encompasses various levels such as image [2], [3], [4], [5], feature [6], [7], [8], and decision [9].

Feature-level fusion integrates and models feature representations from various sources before the final classification step. In the video classification framework, feature-level fusion incorporates both spatial and temporal features from video frames to produce a more complete and discriminative video representation. Common methods for feature-level fusion include concatenation, element-wise summation, weighted averaging, and multi-modal attention mechanisms. The combined features are subsequently fed into a classifier in order to improve classification performance [10], [11], [12].

On the contrary, decision-level fusion combines the predictions of multiple classifiers and models to classify a video. In the context of video action classification scenarios, it is common to train multiple models with different architectures, input data, or hyperparameters. The individual models function autonomously to generate separate predictions for identical video sequences. These distinct outputs generated by the individual models are then consolidated using various methodologies, including voting, averaging, weighted averaging, or more advanced ensemble techniques. The main goal of decision-level fusion is to exploit the diversity and complementary capabilities of individual classifiers, enabling them to compensate for each other's limitations and produce action predictions that are more robust and accurate.

Video classification poses challenges for traditional CNN methods, particularly in capturing long-range dependencies and temporal changes inherent in video data. Researchers have turned to transformers, initially developed for Natural Language Processing (NLP) tasks, to address these challenges [13], [14], [15]. Transformers efficiently integrate spatial and temporal features from video frames, allowing for the modelling of extended temporal dynamics and the identification of complex motion patterns, essential for accurate action recognition.

Notable research gaps include the need for the development of efficient fusion techniques, memory-efficient multi-level fusion approaches, and comparative studies of hybrid fusion techniques in video classification. Key research questions revolve around the impact of integrating spatial and temporal features, the effectiveness of multi-level fusion techniques, and the potential of vision transformer based approaches in video classification compared to ConvLSTM-based approaches. This study aims to investigate how feature fusion and decision fusion affect the final video prediction performance without applying complex deep learning architectures or large labelled datasets. Multiple single and multi-stream neural networks have been developed and compared in terms of the classification accuracy across different fusion levels. Additionally, a novel methodology employing transformers for feature fusion is proposed, and various attention mechanisms are compared for video classification performance. The proposed method effectively captures and incorporates spatio-temporal relationships within video frames, resulting in more robust and context-aware action recognition.

This paper makes significant contributions by conducting a comprehensive analysis of the effects of feature and decision fusion methods on video classification performance across different levels and proposing a novel method using vision transformers to combine features, highlighting their effectiveness in capturing and integrating spatial and temporal relationships in video frames. Moreover, this study compares different attention strategies for vision transformers, providing insights into their impact on the accuracy of video classification.

This paper is organised as follows: related work is reviewed in Section II, the key components of the proposed methods are explained in Section III, experimental results are presented and discussed in Section IV, and finally Section V concludes the paper.

II. RELATED WORK

Video classification has been a challenging research topic owing to the complex and dynamic structure of video data. This section provides an overview of the existing research on video classification, focusing on how the current studies tackle temporal information and the way of combining spatial and temporal features.

The conventional approaches extract user-defined features, including Histograms of Optical Flows (HOF) [16] and Histogram of Oriented Gradients (HOG) [17], which are used as inputs to classifiers such as neural networks for video classification. These pre-computed, feature-based classification approaches are typically constructed as multiple streams to encode spatial and temporal features [18], [19], [20], [21]. The commonly used motion features that have been pre-computed include motion history/energy images [22], [23], [24], dynamic images [25], [26], and joint trajectory maps [27], [28], [29] in multi-stream neural networks. However, a significant challenge is the effective communication between streams to exchange information for the acquisition of multi-modal spatio-temporal representations. Although multi-stream networks use information fusion techniques to effectively combine information from different modalities [30], [31], [32], these techniques may not be sufficient to model long-term dependencies [19], [22], [33]. This makes it challenging to get accurate temporal information at the video level.

A typical machine learning approach uses 3D filters to learn spatio-temporal features from videos using 3D CNNs for video classification. 3D CNNs are the extended version of 2D CNNs with the addition of a temporal dimension to the convolutional filters, where the temporal dynamics of actions are extracted [34]. The third dimension of the CNN is employed to extract motion information between consecutive frames by allowing still images to be ordered systematically [20]. Tran et al. [34] trained deep 3D CNNs on large-scale action datasets and they demonstrated that 3D CNNs are more effective in action recognition for spatio-temporal feature extraction than 2D CNN architectures. Moreover, the effects of different spatio-temporal convolutions have been investigated and 3D CNNs outperformed 2D CNNs in the concept of residual learning [35]. The use of 3D filters in combination with other techniques, such as motion-based features and transformers, has improved the accuracy and efficiency of action recognition [36], [37], [38], [39], [40]. 3D CNNs have the ability to extract distinctive features in spatial and temporal dimensions, requiring simultaneous processing of both types of features. However, there are a larger number of parameters in 3D CNNs, and their computational complexity is much higher than that of 2D CNNs. This increased model complexity results in increased demands for memory resources and thus reduces the available hardware resources for training and developing 3D CNNs.

Another machine learning approach involves the integration of features extracted by CNNs from individual frames with temporal sequence architectures, most commonly known as RNNs. RNNs are purposefully engineered to handle sequential data by leveraging past information within the sequence to produce the classification output [41], [42], [43]. Nevertheless, one prominent concern associated with RNNs pertains to their vulnerability to the short-term memory problem, which is a consequence of the vanishing and exploding gradients problems [44], [45], [46]. The RNN exhibits an increased susceptibility to the issue of vanishing gradients as the number of steps in a sequence that it processes increases, in contrast to alternative neural network architectures. In order to overcome this constraint, researchers have developed specialised variations of RNNs, known as Gated Recurrent Units (GRUs) [47] and Long Short-Term Memory (LSTM) [44] networks. LSTM and GRU architectures are designed to effectively retain information from prior data over extended sequences. This is accomplished by employing gating mechanisms, which enable the memory cells to store and retrieve relevant information. The function of these gates is to regulate the transmission of information within the network, allowing them to acquire knowledge about the importance of inputs in a given sequence and efficiently retain or transmit their information across long sequences. In action recognition, CNN and LSTM are used for extracting spatial and temporal patterns in video sequences [48], [49], [50]. Additionally, ConvLSTM [51], [52], [53] and bidirectional LSTM [54] have been investigated in video classification to access long-distance dependencies over video inputs.

Another method for video classification is based on attention mechanisms and transformers. Transformer models have been introduced in recent years for NLP, including *BERT* [13], *GPT* [14], *RoBERTa* [15], and *T5* [55], which have demonstrated promising results in classification and translation tasks. Consequently, transformers have been integrated into computer vision, a field that has relied heavily on deep ConvNets and RNNs. Popular transformer-based models for computer vision tasks include *ViT* [56] and *DeiT* [57] for image classification, and *VisTR* [58] for video instance segmentation. Action recognition, with its sequential nature of video data, aligns well with the capabilities of transformers for modelling temporal variations. Despite being relatively new in the field of action recognition, transformers have spurred a significant amount of research in recent years [59], [60], [61]. Utilising attention-based mechanisms without the need for RNN backbones, transformers showcase their capacity compared to RNNs in combination with attention. Some approaches solely rely on transformers and self-attention mechanisms for extracting spatio-temporal features, whereas others combine CNN features with transformers, capitalising on the strengths of both architectures [62], [63], [64].

Hybrid techniques in video analysis encompass the combination of various approaches to effectively capture the spatial and temporal dynamics. Illustrative instances encompass the integration of motion-based characteristics with 3D filters or the utilisation of a hybrid approach involving LSTM networks and 3D CNN models. The use of 3D CNN is commonly adopted in hybrid methods, while the incorporation of transformers is a more recent and less extensively studied approach. Researchers seek to take the advantage of distinctive benefits of various temporal modelling techniques by employing them on video sequences. In the motion-based and 3D CNN approach, the utilisation of motion-based features offers a direct method to comprehend the overall motion and the transitions that occur between consecutive frames. The acquisition of these transitions can subsequently appear through the gradual acquisition of 3D filters, facilitating the hierarchical acquisition of motion through separate methodologies. Although hybrid techniques have the potential to offer various advantages, there remains a lack of research investigating their complete capabilities. Furthermore, there is a lack of extensive exploration regarding the most effective methods for integrating these approaches, highlighting the necessity for additional investigation and experimentation in this domain [40].

III. METHODS

This section describes the design and evaluation of the methods proposed for video classification in this study, including the deep learning architectures, evaluation methodologies, different information fusion mechanisms, and vision transformer based approach to feature fusion. The focus of the proposed methods is on improving video classification performance through strategic feature and decision fusion techniques.

Our methodology is centred around the exploration of novel techniques for optimising video classification performance. The proposed method for fusing features and decisions across multiple levels can strategically capture extensive temporal dependencies between video frames. This paper also proposes to use vision transformers for effective feature fusion, aiming to capture intricate spatio-temporal relationships within video frames. This novel method is designed to enhance context awareness by integrating spatial and temporal information effectively. Integrating this vision transformer based approach alongside existing deep learning architectures and fusion mechanisms leads to a promising framework for video classification.

A. NETWORK ARCHITECTURES

Multiple CNN-based networks integrating single-stream and multi-stream CNNs are developed in this study. The rationale behind the selection of these architectures lies in their adaptability to capture both spatial and temporal features from diverse data sources, enhancing the model's ability to comprehend complex dynamic video data.

For the single-stream design, inspired by the ConvLSTM architecture used in our earlier research, the incorporation of RGB frames *(RGB)*, HOF features *(HOF)*, and HOG features *(HOG)* aims to leverage specific feature sets for comprehensive information extraction.

Regarding the CNN-based architectures, the utilisation of ConvLSTM as the cornerstone for establishing a baseline video classification framework was motivated by its proficiency in encapsulating spatio-temporal dynamics within time series data. As shown in Fig. 1, a ConvLSTM layer is included onto the spatial feature maps derived from VGG-16 pre-trained network, and the resulting hidden states are utilised for the purpose of classification. In order to conduct the experiments for ConvLSTM-based architectures, ConvLSTM layer contains 64 hidden states, 7×7 kernels, 0.2 dropout, and a convolution stride of 2×2. The number of units in the output space was set to 1024 and ReLU was used as the activation function. A 0.6 dropout was also applied before the final dense layer.

In this study, information fusion has been applied at two levels. At the feature level, several sets of features are extracted from each video, including motion features and appearance features. These features are then merged using a feature-level fusion technique to create a single feature vector for each video. Moreover, at the decision level, multiple classifiers are used, each trained on a different feature set. The outputs of the classifiers are then combined using a decision-level fusion technique to produce the final classification result.

To investigate the impact of various information fusion levels on video classification performance, several deep multi-stream neural networks were implemented. Similarly for this purpose, a multi-level fusion approach was proposed by merging both feature and decision fusions in multi-stream neural networks. Furthermore, for the feature fusion, a novel vision transformer based multi-stream neural network architecture was proposed and an ablation study was conducted to evaluate the effect of various attention mechanisms in the transformer encoder.

B. DECISION FUSION

Two types of classification architectures are developed in this study: single-level and multi-level. Different features are fed into the classifiers in each stream of the single-level architecture, as shown in Figs. 2 and 3. The classifiers are trained independently and the ultimate decision is made depending on the average prediction scores from those classifiers. As shown in Fig. 2, two-stream architectures have two distinct inputs: (1) RGB frames and HOF representations (*RGB-HOF*), (2) RGB and HOG features (*RGB-HOG*), or (3) HOG features and HOF representations (*HOG-HOF*). As can be seen from Fig. 3, a three-stream neural network was implemented to assess classification performance using all three input formats (*RGB-HOG-HOF*). Furthermore, as demonstrated in Fig. 4, a novel multi-level decision approach has been applied for the final decision to be made progressively based on the scores achieved from each pair of inputs. The network designs used in this study are listed in Table 1, where both spatial and temporal representations are combined in the proposed single-level and multi-level classifiers.

The softmax activation function is used in the output layer of the proposed neural networks to observe the categorical distribution of a given set of labels and to compute the probability that each input component corresponds to a specific class. The softmax function $S : \mathbb{R}^K \rightarrow \mathbb{R}^K$ is defined by the following equations:

$$S_i(x) = \frac{exp(x_i)}{\Sigma_j^K exp(x_j)} \tag{1}$$

where $i = 1, \ldots, K$ and $x = [x_1, \ldots, x_K]$ is the output of the final dense layer of the neural network. In this study, the softmax scores from the network streams are averaged to give a better representation than any single stream could provide.

TABLE 1. Decision fusion based architectures used in the experiments.

Architecture	Fusion Level	Input
D-RGB-HOF	single-level	RGB-HOF
D-RGB-HOG	single-level	RGB-HOG
D-HOG-HOF	single-level	HOG-HOF
D-RGB-HOG-HOF	single-level	RGB-HOG-HOF
D1	multi-level	D-RGB-HOF and D-RGB-HOG
D2	multi-level	D-RGB-HOF and D-HOG-HOF
D3	multi-level	D-RGB-HOG and D-HOG-HOF

C. FEATURE FUSION

Similar to the decision-level fusion approach, single-level and multi-level architectures are designed in this study for feature fusion. In single-level feature fusion, low-level feature frames are individually inputted into pre-trained VGG-16 neural networks. The acquired high-level features from VGG-16 are then concatenated as the input vector of

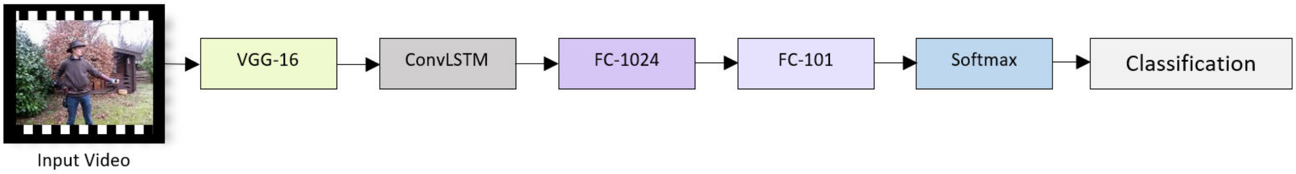

FIGURE 1. The ConvLSTM video classification architecture employed in the experiments.

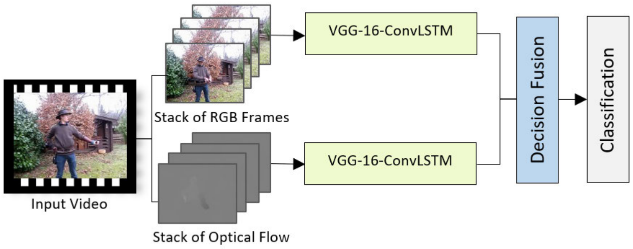

FIGURE 2. Illustration of the two-stream ConvLSTM architecture employed in the experiments involving RGB frames and optical flows. Additionally, single-level fusion architectures incorporating alternative feature combinations were designed, as detailed in Table 1.

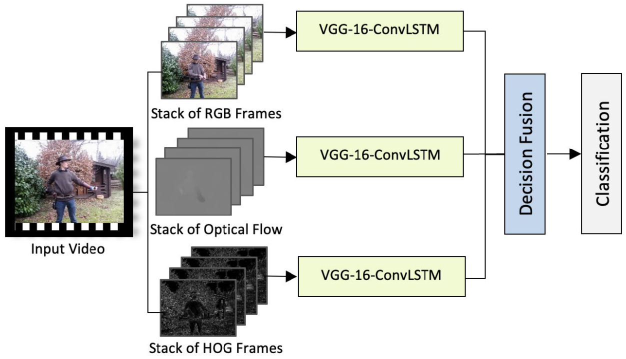

FIGURE 3. Illustration of the single-level decision fusion architecture employed in the experiments utilising RGB frames, optical flows, and HOG features.

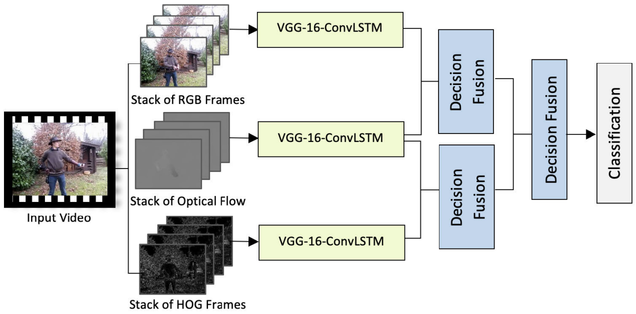

FIGURE 4. Illustration of the multi-level decision fusion architecture employed in the experiments involving RGB frames, optical flows, and HOG features. Additionally, multi-level fusion architectures incorporating alternative feature combinations were designed, as detailed in Table 1.

TABLE 2. Feature fusion based architectures used in the experiments.

Architecture	Fusion Level	Input
F-RGB-HOF	single-level	RGB-HOF
F-RGB-HOG	single-level	RGB-HOG
F-HOG-HOF	single-level	HOG-HOF
F-RGB-HOG-HOF	single-level	RGB-HOG-HOF
FD1	multi-level	F-RGB-HOF and F-RGB-HOG
FD2	multi-level	F-RGB-HOF and F-HOG-HOF
FD3	multi-level	F-RGB-HOG and F-HOG-HOF

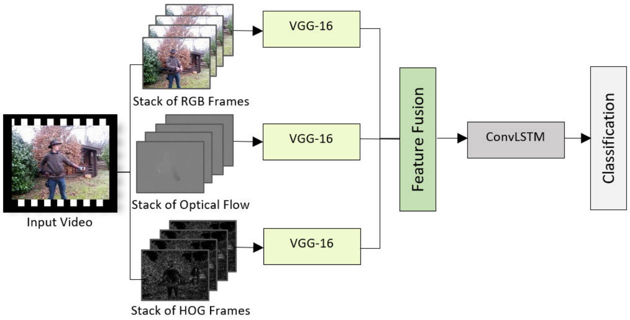

FIGURE 5. Illustration of the single-level feature fusion architecture employed in the experiments involving RGB frames, optical flows, and HOG features. Additionally, single-level fusion architectures incorporating two input channels were designed, as detailed in Table 2.

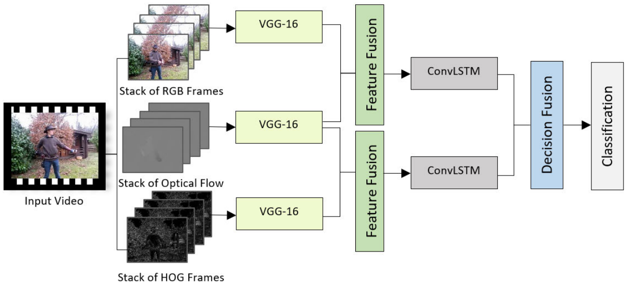

FIGURE 6. Illustration of the multi-stream fusion architecture employed in the experiments combining multi-level feature and decision fusion. Additionally, multi-level feature and decision fusion architectures incorporating alternative combinations were designed, as detailed in Table 2.

the ConvLSTM final classifier, as depicted in Fig. 5. The experiments have investigated multiple single-level feature fusion architectures incorporating different combinations of feature types, as listed in Table 2.

In the experiments, the effect of fusing features and decisions in multi-stream neural networks for video classification was also investigated. Spatial and temporal features are combined by using feature fusion and then the softmax scores are fused based on the average rule as shown in Fig. 6. Three different architectures for combining feature fusion and decision fusion were implemented and tested, *FD1*, *FD2*, and *FD3*, as listed in Table 2.

D. VISION TRANSFORMER BASED FUSION

A novel vision transformer based fusion architecture is proposed in this study to classify videos by effectively combining spatial and temporal information. As shown in Fig. 7, the input of the model comprises the integration of RGB frames, HOF and HOG features, which capture appearance, motion, and spatial representations, respectively. The proposed architecture based on vision transformers is implemented to process the multi-modal input in order to make predictions regarding the corresponding action category.

The encoder-decoder structure, similar to that in *VIVIT* [62] is adopted in our vision transformer based fusion architecture. The encoder is responsible for receiving the sequence of input, whereas the decoder is responsible for generating the output sequence.

The input sequence in the transformer is divided into three distinct components, namely queries, keys, and values, which are fed to the transformer encoder. The encoder is compromised of a series of identical layers, where each layer is composed of multiple sub-modules. These sub-modules include normalisation layers, multi-head attention mechanisms, and Feed-Forward Neural Networks (FFNNs), as depicted in Fig. 7. The encoder receives a series of tokens as its input, which are then processed subject to the attention mechanism. This attention mechanism allows the encoder to focus on relevant data from various positions within the sequence. The architectural design in the VIVIT involves self-attention mechanisms in order to effectively capture contextual information present in both the input and output sequences.

In this paper, RGB, HOF, and HOG features are fused first for mapping a video to a sequence of tokens as input of the proposed transformer-based fusion architecture. Following that, positional tubelet embeddings are adopted, as described in [62], in order to acquire non-overlapping spatio-temporal input patches for the transformers. The embedding approach involves the extraction of tokens from the temporal, height, and width dimensions. This method also incorporates spatio-temporal information during tokenization, in an intuitive manner.

This study also investigates the efficacy of the transformer-based fusion approach by employing different attention mechanisms, including *Bahdanau's Dot-Product Attention* [65], *Luong's Additive Attention* [66], and *Multi-head Self Attention* [67].

Bahdanau's Dot-Product Attention mechanism calculates the significance of each position in the input sequence in relation to the existing decoding process. It employs a trainable alignment model, FFNN, to produce attention weights between the queries, which are subsequently employed to calculate the context vector by performing a weighted summation of the encoder's output [65], as in Eq. (2):

$$c_t = \sum_{i=1}^{T_x} a_{t,i} h_i \tag{2}$$

where each hidden state h_i is weighted by $a_{t,i}$. The weight for each hidden state is also aligned by the softmax function defined in Eq. (1). The attention score is calculated as in Eq. (3):

$$score(s_t, h_i) = v_a^T \tanh(W_a[s_t; h_i]) \tag{3}$$

where v_a and W_a are weight matrices learned by alignment model when h and s are hidden states of the encoder and decoder, respectively.

Luong's Additive Attention mechanism calculates the correspondence between the hidden state of the decoder at a given time step and the output of the encoder, like the *Bahdanau's* attention. Nevertheless, it applies additive operations for computing attention weights, which are more simple and straightforward than *Bahdanau's* attention where a neural network is used for it [66]. The attention score in *Luong-sytle* attention mechanism is calculated as follows:

$$score(s_t, h_i) = s_t^T h_i \tag{4}$$

Bahdanau's and *Luong's* attention mechanisms are global approaches as all the input tokens and all hidden states are considered in the context vector. Although they can remember the sequential input, this process is complex, computationally expensive, and improper for long sequences. Moreover, in these single-head attention methods, the computation of a set of key-value pairs is performed on an input sequence, given a query. The determination of the attention weight between the query and each key is accomplished by employing compatibility functions (like Eqs. (2) and (3)). The weights assigned to each position in the input sequence indicate the level of relevance or significance in relation to the provided single query, thus one representation can be learnt.

The *Multi-Head Attention* mechanism plays an important role in the transformer architecture, as it aims to effectively capture contextual information and long-range dependencies within sequences [67]. In this mechanism, prior to calculating the attention weights, the input sequence goes through a process of converting it into query (Q), key (K), and value (V) representations through learnable linear transformations. This attention method enables the model to simultaneously focus on different parts of the input sequence, utilising multiple attention heads. Each attention head is regarded as an independent attention mechanism that possesses its own set of query, key, and value transformations that can be learned. This enables the model to simultaneously focus on various elements of the input sequence, thereby improving its capacity to comprehend a wide range of patterns and connections. After the computation of attention weights for each head, the outputs of the multiple heads are concatenated and subjected to a linear transformation using an additional trainable linear layer. The final output of the *Multi-Head Attention* mechanism is generated by aggregating the information from all the heads as follows:

$$Attention(Q, K, V) = concat(head_1, \ldots, head_h)W^O \tag{5}$$

where $head_i = Attention(QW_i^Q, K\ W_i^K, V\ W_i^V)$.

In transformer-based fusion experiments, the input data is structured as a tensor with dimensions of 60×7×7×512,

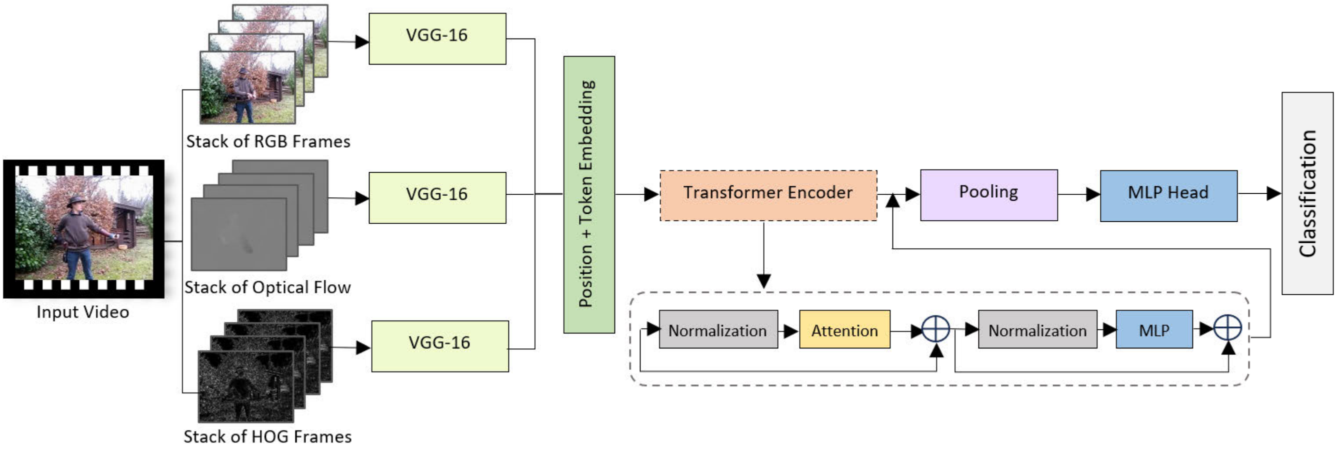

FIGURE 7. Illustration of the multi-stream vision transformer based fusion network architecture employed in the experiments combining RGB frames, HOF, and HOG features.

obtained by combining RGB, HOG and HOF features. The size of the tubelet embedding path is 7×7×7. The Transformer model used in our study consists of a total of 8 layers. In multi-head attention, we use 20 heads and the key dimension in the multi-head attention is determined by dividing the embedding size by the number of heads. In the transformer model, two densely connected layers with GELU activation functions are employed, with 2048 units in the first layer and 512 units in the second layer. Just before the final softmax layer, global average pooling is applied to extract essential global information from the feature maps. This comprehensive architecture ensures the model's ability to effectively capture and classify video data across various spatio-temporal features and frames.

To comprehensively assess the impact of each attention mechanism, an ablation study was conducted. This study systematically evaluates the model's performance using different combinations of attention mechanisms. The primary goal of this investigation is to identify the most optimal information fusion approach that can yield high accuracy in video classification tasks.

E. DATASETS AND EXPERIMENTAL SETUP

In this work, the KTH and UCF-101 datasets, which are widely recognised benchmark datasets for human action detection from videos, are used for performance evaluation. The KTH dataset consists of six activity categories: walking, jogging, running, boxing, hand waving, and hand clapping. The UCF-101 dataset includes videos of individuals engaging in a variety of activities, such as playing instruments, engaging in sports, and interacting with objects. The KTH dataset was selected for initial experiments because it is considered as a small-sized dataset and the experiments were repeated on the UCF-101 dataset, which gives the largest diversity in human actions with its 101 action categories. Both datasets have gained popularity within the computer vision community due to their ability to provide a complex problem for action recognition algorithms.

In the KTH dataset [68], the videos are recorded in controlled environments and staged several times by 25 different actors with different clothes in four outdoor scenarios. The video clips were captured over static backgrounds with a consistent frame rate of 25 frames per second (FPS) at a size of 160×120 pixels. The number of classes is very low in comparison to the extensive range of human actions in real life. The number of classes plays an important role in the assessment of an action recognition method [4, 9]. Therefore, we conducted tests using the UCF-101 dataset to further investigate this matter.

The UCF-101 dataset [69] comprises web videos that are captured in uncontrolled settings, often exhibiting camera movement, diverse lighting conditions, partial obstruction, and low-quality frames. The videos were obtained from the online platform YouTube, and any samples considered redundant were manually eliminated. Each video clip has a dynamic background and a fixed frame rate of 25 FPS with a resolution of 320×240 pixels.

Several data preprocessing and transformation procedures were executed on both datasets in order to facilitate video classification. These steps are crucial to ensure that the data is in a suitable format for training and evaluating deep learning models. Initially, the original datasets were partitioned into training and test subsets based on their officially designated benchmarking partitions. Additionally, a validation subset was set aside for the purpose of tuning hyperparameters during the model development process. The video clips were converted into individual frames and the frames were resized to a consistent size (224×224 pixels) to maintain uniformity across all samples. During this stage, we extracted and stored VGG-16 features, HOG and optical flows in separate directories. In order to effectively manage the large volumes of input data for deep models, the utilisation of data batches was employed. Additionally, a data batch generator was implemented to facilitate the efficient loading and preprocessing of the data, hence enabling parallel processing. In the process of label encoding, the textual representation of each class was transformed into a numerical format.

In order to conduct a comprehensive evaluation of the implemented classification architectures on the datasets, many experimental iterations have been carried out, each incorporating a distinct combination as outlined in previous sections. The results obtained from the validation set are used for the purpose of tuning the hyperparameters.

For the hyperparameter tuning, a manual search approach is adopted, guided by insights from prior research endeavours. In this approach, various sets of hyperparameters were established using subjective evaluations or prior knowledge. The deep architectures have subsequently been trained using these sets, their performance has been assessed, the training process iteratively refined until a desirable level of accuracy was attained, and ultimately the optimal set of hyperparameters that yielded the highest accuracy was identified. The initial learning rate was set to 0.003. This rate was chosen based on preliminary experiments and empirical observations to ensure stable training. A batch size of 32 was employed during the training of the model. The choice of batch size was made in order to strike a compromise between computing efficiency and model convergence for both datasets.

The validation of models is an essential component of our training approach, especially when dealing with extensive datasets that necessitate the loading and processing of data in batches. To ensure the effectiveness of model validation, validation was performed on the entire validation dataset at the end of each epoch. The utilisation of this methodology facilitated the execution of a comprehensive assessment of the model's performance.

To prevent overfitting and optimise model performance, early stopping was implemented during the training process. Early stopping was determined based on the validation loss, monitoring the loss on the validation dataset at the end of each epoch. The patience value was set to 10, meaning that if the validation loss did not improve for 10 consecutive epochs, training was terminated to prevent overfitting. Data shuffling was applied to prevent the model from overfitting to specific patterns in each epoch, thereby improving generalisation. Dropout layers were also incorporated in both the transformers-based and ConvLSTM-based models. The final selected models are evaluated on the test set in order to mitigate any potential biases arising from overfitting.

F. PERFORMANCE EVALUATION

The accuracy score was employed as a main performance indicator in our video classification tests to evaluate the efficacy of our fusion approaches and benchmark models. Accuracy is a crucial parameter in the context of video classification since it quantifies the ratio of accurately categorised video samples to the overall number of samples. Furthermore, the confusion matrix was employed in our experiments, serving as a tabular representation that offers a comprehensive breakdown of true positives, true negatives, false positives, and false negatives. This visual representation allows for an assessment of the model's performance across various classes, facilitating the identification of specific areas in need of development. The evaluation of model confusion was conducted throughout the process of model development and model selection.

Our study includes two different evaluation processes on the gathered test accuracy scores: (1) evaluating single and multi-stream neural network architectures in terms of feature and decision fusion in single and multi-levels, (2) evaluating vision transformer based feature fusion approaches based on different attention mechanisms.

For the first evaluation, using the *Kolmogorov-Smirnov* [70] test, the observed cumulative distribution was compared to the cumulative distribution that would be expected if the data were normally distributed. According to the *Levene* statistic [71], group variances are homogeneous based on the mean and median. The *ANOVA* test was also used to compare variance differences and determine the significance of the results. The *Tukey's HSD* test was used to determine whether the group means are distinct.

For the second evaluation process, the classification performances were compared using the *Independent T-test* to figure out whether the difference between a pair of attention mechanisms is significant. This analysis helps identify whether any attention mechanism exhibits superior performance compared to the others.

Through the use of these extensive evaluation methods, this study not only evaluates the performance of the proposed methods for video classification but also establishes a direct link between quantitative evaluation indicators and their practical implications in real-world situations. The objective of this approach is to enhance the applicability and comprehension of our findings beyond conventional metrics.

IV. RESULTS AND DISCUSSION

This section presents the comprehensive results obtained from the experiments. The objective of this study is to assess and compare the performance of different ConvLSTM-based single and multi-level architectures for decision and feature fusion, along with feature fusion using vision transformer based classification architectures. Furthermore, an ablation study was undertaken to evaluate different attention mechanisms employed in the transformer-based architectures.

A. COMPARATIVE ANALYSIS OF ARCHITECTURAL PERFORMANCE

In this part of the study, a total of 17 ConvLTSM-based decision and feature fusion architectures were investigated. On the KTH dataset, 12 separate experiments were executed for each architecture (a total of 204 runs), whereas

15 experiments were run on the UCF-101 dataset (a total of 255 trials). Accuracy scores for training, validation, and testing were recorded after each run, and the results were then analysed and discussed.

Table 3 shows the average accuracy scores obtained by single-stream neural networks, with *RGB*, *HOG*, or *HOF* as input respectively. The *HOF* architecture scored the highest for test accuracy, 81.44% and 73.66% on the KTH and UCF-101, respectively. The crucial importance of motion information in video classification is demonstrated by the temporal network, *HOF*, which performed noticeably better than the spatial networks, *RGB* and *HOG*. On the KTH dataset, the *HOG* architecture outperformed the *RGB* architecture by about 4%, whereas on the UCF dataset, its performance was relatively close to the performance of the *RGB* architecture. This discrepancy may be caused by spatial differences in the two datasets, including those in the number of colour channels, the intricacy of the actions, the intensity of the colours, and other dynamics associated with appearance.

TABLE 3. Accuracy scores obtained by single-stream networks on the KTH and UCF-101 datasets during training, validation, and testing.

Architecture	Dataset	Training Accuracy	Validation Accuracy	Testing Accuracy
RGB	KTH	87.18%	65.89%	73.07%
HOG	KTH	92.40%	70.62%	77.93%
HOF	KTH	98.52%	73.41%	81.44%
RGB	UCF	86.56%	68.77%	70.03%
HOG	UCF	88.33%	68.21%	70.80%
HOF	UCF	93.81%	70.64%	73.66%

Table 4 provides a summary of the multi-stream neural network architectures' performances. Single-level feature fusion based architectures generally outperformed decision fusion based approaches. The architectures fed with optical flows and HOG features, *D-HOG-HOF*, *D2*, *F-HOG-HOF*, outperformed those with other feature pairings in both designs. However, combining *RGB*, *HOG*, and *HOF* data at the feature level produced the best classification results, 84.10% on the KTH and 77.98% on the UCF-101. Fusion architecture *D2* is the second best when both RGB and HOG features are predominately integrated with HOF features. These findings support the notion that temporal information is crucial for video classification.

TABLE 4. Test accuracy scores obtained by multi-stream networks on the KTH and UCF-101 datasets.

Architecture	Fusion Level	KTH Test Accuracy	UCF Test Accuracy
D-RGB-HOG	single-level	78.24%	71.89%
D-RGB-HOF	single-level	81.83%	74.87%
D-HOG-HOF	single-level	82.91%	76.32%
D-RGB-HOG-HOF	single-level	81.79%	75.89%
D1	multi-level	79.78%	73.39%
D2	multi-level	83.68%	76.31%
D3	multi-level	80.90%	74.23%
F-RGB-HOG	single-level	71.80%	73.34%
F-RGB-HOF	single-level	81.25%	74.89%
F-HOG-HOF	single-level	81.33%	75.79%
F-RGB-HOG-HOF	single-level	84.10%	77.98%
FD1	multi-level	80.75%	73.40%
FD2	multi-level	81.48%	75.16%
FD3	multi-level	82.25%	76.07%

TABLE 5. One-way between-groups ANOVA of classification performances of different fusion architectures on the KTH and UCF-101 datasets, where *df*, *SS*, *MS* and *F* refer to degrees of freedom, sum of squares, mean sum of squares, and F score, respectively.

		SS	df	MS	F	Sig.
KTH	Between Groups	.217	16	.014	85.239	.001
	Within Groups	.030	187	.000		
	Total	.247	203			
UCF-101	Between Groups	.105	16	.007	209.524	.001
	Within Groups	.007	238	.000		
	Total	.113	254			

The test accuracy scores achieved by different fusion architectures on the KTH and UCF-101 datasets were statistically different at the $p < 0.05$ level, as shown in Table 5: F (16, 187) = 85.239; p = 0.001 and F (16, 238) = 209.524; p = 0.001, respectively. Tukey's HSD test was used to make additional comparisons. The findings showed that across the majority of the fusion architectures, there were significant differences in classification performance.

Tables 6 and 7 list the architectural groups whose average test accuracy scores on the KTH and UCF-101 datasets were significantly different. The homogeneous subset findings from the post hoc test are provided with the groups in increasing order of average accuracy scores, where the architectures in each group are not statistically different from each other. For instance, in Group 1, *F-RGB-HOG* and *RGB*, achieved the lowest average accuracy, whereas in Group 8, *D-HOG-HOF*, *D2*, and *F-RGB-HOG-HOF*, achieved the highest average accuracy on the KTH dataset, as shown in Table 6. The fusion architectures were divided into 10 groups in terms of statistically significant performance difference on the UCF-101 dataset, compared to 8 groups on the KTH dataset. Due to the fact that *F-RGB-HOG-HOF* does not occur in a subset with any of the other groups on the two datasets, this feature fusion architecture is significantly different from all other groups and architectures.

The proposed multi-level decision and feature fusion methods, *D2* and *FD3*, have yielded results that are comparable to those of the *F-RGB-HOG-HOF* method and have the advantage of requiring lower memory and computational cost, which implies that higher-dimensional feature vectors are advantageous but computationally expensive. Thus applying dimensionality reduction techniques can be effective for this method. By optimising the fusion algorithm by employing other weighting or fusion strategies, the accuracy scores obtained by the proposed methods can also be further improved. However, conducting further experiments requires a substantial amount of resources.

Multi-level architectures *FD1*, *FD2*, and *FD3*, which integrate decision and feature fusion, achieved results comparable to those from feature fusion or decision fusion. On the KTH dataset, no significant difference was observed among these approaches. However, on the UCF-101 dataset, there is a significant difference between all possible combinations.

The above experimental findings clearly show that, across all evaluation metrics, feature-level fusion consistently outperformed decision-level fusion. By utilising both spatial and temporal information, the feature fusion strategy showed superior discriminative capabilities, which improved classification accuracy. Therefore, feature fusion methods were

TABLE 6. One-way between-groups ANOVA of classification performances obtained by different fusion architectures on the KTH dataset.

Architecture	1	2	3	4	5	6	7	8
F-RGB-HOG	71.80%							
RGB	73.07%							
HOG		77.93%						
D-RGB-HOG		78.24%	78.24%					
D1			79.78%	79.78%				
FD1				80.75%	80.75%			
D3				80.90%	80.90%			
F-RGB-HOF				81.25%	81.25%	81.25%		
F-HOG-HOF				81.33%	81.33%	81.33%		
HOF				81.44%	81.44%	81.44%		
FD2				81.48%	81.48%	81.48%		
D-RGB-HOG-HOF					81.79%	81.79%		
D-RGB-HOF					81.83%	81.83%		
FD3					82.25%	82.25%	82.25%	
D-HOG-HOF						82.91%	82.91%	82.91%
D2							83.68%	83.68%
F-RGB-HOG-HOF								84.10%
Sig.	0.527	1.000	0.197	0.092	0.233	0.113	0.318	0.638

TABLE 7. One-way between-groups ANOVA of classification performance obtained by different fusion architectures on the UCF-101 dataset.

Architecture	1	2	3	4	5	6	7	8	9	10
RGB	70.03%									
HOG		70.80%								
D-RGB-HOG			71.89%							
F-RGB-HOG				73.34%						
D1				73.39%						
FD1				73.40%						
HOF				73.66%	73.66%					
D3					74.23%	74.23%				
D-RGB-HOF						74.87%	74.87%			
F-RGB-HOF						74.89%	74.89%			
FD2							75.16%	75.16%		
F-HOG-HOF								75.79%	75.79%	
D-RGB-HOG-HOF									75.89%	
FD3									76.07%	
D2									76.31%	
D-HOG-HOF									76.32%	
F-RGB-HOG-HOF										77.98%
Sig.	1.000	1.000	1.000	0.980	0.296	0.115	0.991	0.157	0.449	1.000

further investigated in the remaining experiments, especially focusing on the vision transformer based feature fusion method.

B. ABLATION STUDY ON ATTENTION MECHANISMS

Recently, vision transformer based classification architectures have made significant contributions in the context of integrating features for classification tasks. Their advantages stem from their intrinsic capacity to capture dependencies over long distances, model temporal relationships over time, and acquire intricate patterns that exist in video data. In the remaining experiments, spatio-temporal transformer based architectures were used on top of the pre-extracted features to test their effect on classification performance improvement. In particular, the effect of different attention mechanisms used in spatio-temporal transformers were investigated.

The performance of transformer-based fusion architectures was assessed by comparing them against the best ConvLSTM-based feature fusion architecture, i.e., *F-RGB-HOG-HOF*, through a comparative analysis. Table 8 shows that transformer-based architectures significantly outperformed the ConvLSTM-based feature fusion architectures on both KTH and UCF-101 datasets.

According to Tables 9 and 10, which show the T-test results, the video classification performance was significantly improved by all transformer-based architectures over *F-RGB-HOG-HOF*. On both KTH and UCF-101 datasets and across different evaluation metrics, the transformer-based classification architectures achieved superior performance. The p-values are provided for comparison, serving as an indication of the statistical significance of the observed variations in accuracy. All calculated p-values are found to be lower than the predetermined threshold of 0.05, indicating statistically significant differences in performance among the tested pair of fusion methods. The results confirm the consistency and reliability of the observed enhancements in the classification performance achieved by the classification architecture based on transformers in comparison to the feature fusion based on ConvLSTM. This finding highlights the capacity of transformer models to proficiently capture intricate patterns and relationships within the data, a critical aspect for attaining higher accuracy in classification tasks.

In order to obtain a more comprehensive understanding of the performance differences observed across various fusion techniques, an ablation study was also conducted, which employed a systematic approach to deactivate a specific attention mechanism inside the transformer encoder of the fusion methods, with the subsequent evaluation of their respective effects on classification accuracy. The findings from the conducted ablation study demonstrate that *Luong's Additive Attention* mechanism consistently exhibited the highest level of accuracy on both datasets when compared to the other attention mechanisms that were examined in this study (KTH:88% and UCF-101:81%). *Bahdanau's Dot-Product Attention* exhibited the next highest level of accuracy (KTH: 87% and UCF-101: 81%), only slightly worse than Luong's attention. Multi-head attention achieved the lowest accuracy scores when compared to other attention mechanisms that were evaluated (KTH:86% and UCF-101:80%). Although multi-head attention achieved significant performance improvement over ConvLSTM-based feature fusion, it seemed to be less proficient in capturing complex relationships within the data when compared to other attention mechanisms.

TABLE 8. Classification accuracy achieved by feature fusion networks.

Architecture	KTH Test Accuracy	UCF Test Accuracy
F-RGB-HOG-HOF	84.10%	77.98%
Multi-head Self Attention	86.30%	79.98%
Bahdanau's Dot-Product Attention	87.18%	80.52%
Loung's Additive Attention	87.82%	80.73%

C. DISCUSSION

In this study, the interaction between the selection of neural network architecture and the classification performance was investigated for video classification. Different types of neural networks have been developed, where various feature and decision fusion techniques are applied at different levels. Therefore, our study focused on examining the effects of multi-level decision and feature fusion and different attention mechanisms in transformer-based fusion, including self-attention, additive attention, and dot-product attention, on improving classification ability. In this section, we discuss the results of the proposed methods and compare them with the state-of-the-art, considering their advantages and disadvantages in two groups: multi-level feature and decision fusion and transformer-based fusion.

1) MULTI-LEVEL FEATURE AND DECISION FUSION

According to the reported results, on RGB input settings, compared with CNN-based approaches, our RGB and HOG architectures outperform deep networks [18] and spatial stream networks [18]. Our proposed architectures achieved around %80 classification accuracy on the UCF-101 dataset, which is comparable to well-known two-stream temporal stream networks [19], [48], [49]. One of the best-performing methods on the UCF-101 dataset was proposed by Feichtenhofer et al. [72], which uses ResNet-50 models on RGB and HOF streams and achieved 94.6% when combined with the dense trajectories model. Although the use of dense trajectories can increase the classification performance, calculating dense trajectories involves tracking features in multiple scales and orientations, leading to a high computational cost, and this can make the method impractical for real-time or large-scale video analysis applications.

Indeed, the utilisation of extensive, labelled datasets for pre-training significantly contributes to the enhanced prediction performance of deep networks, i.e., Kinetics and Sports 1M pre-trained networks in action recognition [34], [73], [74]. Notably, the current state-of-the-art performance in this field has been demonstrated by Gowda et al. [74]. The methodology employed in their study involves the integration of many methodologies, such as the SMART frame selection method in conjunction with Wang's Temporal Segment Network [75]. The combination presented in their study incorporates data from many modalities, including RGB, HOF, and warped flow, which have been pre-trained using the Kinetics dataset.

The aim of our study is to utilise spatial-temporal information in fundamental neural network architectures without the need for complex structures or extra training data, while also ensuring efficient use of resources. The findings obtained from this study indicate that the multi-level fusion technique did not yield the best performance, but the proposed fusion approach demands fewer computational and memory resources. It can be argued that our method can achieve competitive performance under similar training and testing conditions.

Failing to achieve the best performance by the multi-level fusion approach can be attributed to several factors that require further investigation. In the conducted experiments, all features were treated equally, and the fusion process relied only on concatenation and averaging techniques. This basic methodology employed may not have effectively encompassed the complex nature and details of the underlying dataset. The validity of the assumption that all features have an equal impact on the ultimate outcome may not be universally applicable. It is conceivable that certain features possess greater informativeness or relevance than others, and their equitable treatment may have diminished their impact. For example, the findings of our study show that using optical flows in both streams on multi-stream networks can contribute to reaching higher classification performance.

Moreover, it has been observed that the utilisation of confusion matrices has demonstrated the efficacy of optical flow features in mitigating the ambiguity that arises from the similarity between different action categories in spatial feature based methodologies. On the KTH dataset, it was observed that action pairs with similar spatial features, such as hand waving and hand clapping, running and jogging,

TABLE 9. T-test analysis of classification performance obtained by different feature fusion architectures on the KTH dataset.

Architecture	Mean	Std.	Std. Error	t	df	p
F-RGB-HOG-HOF	0.84	0.005	0.001	-8,870	20	0.000
Multi-head Self Attention	0.86	0.007	0.002			
F-RGB-HOG-HOF	0.84	0.005	0.001	-15,203	20	0.000
Bahdanau's Dot-Product Attention	0.87	0.004	0.001			
F-RGB-HOG-HOF	0.84	0.005	0.001	-23,259	15,854	0.000
Loung's Additivelala Attention	0.88	0.002	0.001			
Multi-head Self Attention	0.86	0.007	0.002	-3,502	18	0.003
Bahdanau's Dot-Product Attention	0.87	0.004	0.001			
Multi-head Self Attention	0.86	0.007	0.002	-6,915	11,028	0.000
Bahdanau's Dot-Product Attention	0.87	0.004	0.001			
Multi-head Self Attention	0.86	0.007	0.002	-6,915	11,028	0.000
Loung's Additive Attention	0.88	0.002	0.001			
Bahdanau's Dot-Product Attention	0.87	0.004	0.001	-4,159	13,373	0.001
Loung's Additive Attention	0.88	0.002	0.001			

TABLE 10. T-test analysis of classification performance obtained by different feature fusion architectures on the UCF-101 dataset.

Architecture	Mean	Std.	Std. Error	t	df	p
F-RGB-HOG-HOF	0.78	0.006	0.001	-10,805	22,902	0.000
Multi-head Self Attention	0.80	0.004	0.001			
F-RGB-HOG-HOF	0.78	0.006	0.001	-13,957	22,716	0.000
Bahdanau's Dot-Product Attention	0.81	0.003	0.001			
F-RGB-HOG-HOF	0.78	0.006	0.001	-17,928	16,056	0.000
Loung's Additive Attention	0.81	0.001	0.000			
Multi-head Self Attention	0.80	0.004	0.001	-3,519	18	0.002
Bahdanau's Dot-Product Attention	0.81	0.003	0.001			
Multi-head Self Attention	0.80	0.004	0.001	-6,386	18	0.000
Loung's Additive Attention	0.81	0.001	0.000			
Bahdanau's Dot-Product Attention	0.81	0.003	0.001	-1,908	11,625	0.081
Loung's Additive Attention	0.81	0.001	0.000			

and running and walking, demonstrated reduced confusion when optical flow features were utilised. The use of temporal changes helped to differentiate action pairs through the extraction of optical flow. Therefore, giving higher weight to the optical flow features would improve the performance further.

Additionally, it is possible that the selection of fusion techniques, such as concatenation and averaging, was not optimal in relation to the characteristics of the datasets or the problem under investigation. Various data types and their respective characteristics may necessitate the utilisation of more advanced fusion techniques that take into account the complex relationships between features and decisions.

Another aspect to take into account is the potential interaction between various features and decisions. Some methods failed to leverage potential synergies between specific features or decisions may be because of the equal treatment of all features and uniform combination of all decisions. Certain combinations of features and decisions may exhibit greater complementarity, resulting in enhanced performance, whereas other combinations may introduce undesirable noise or redundancy.

2) VISION TRANSFORMER BASED FUSION

In the context of video classification problems, the integration of features using transformers holds great potential for enhancing classification performance. Experimental results have shown that multi-stream transformer fusion, particularly through the additive attention mechanism, outperformed the other proposed fusion architectures examined in this study.

Transformer-based networks, widely recognised for their achievements in various tasks related to natural language processing and computer vision, utilise self-attention mechanisms. These techniques play a crucial role in capturing interdependence among features across temporal sequences, allowing the model to selectively extract important information and construct meaningful associations.

On the other hand, the advantage of self-attention is its ability to effectively capture and represent temporal dependencies within a single modality. It is also effective in capturing extensive temporal connections across various time steps. However, there is a potential limitation in its ability to effectively capture complex interactions among diverse modalities, similar to the dot-product attention mechanism.

In our experimental findings, it was shown that both the additive and dot-product attention mechanisms exhibited superior performance compared to the multi-head self-attention mechanism. This observation highlights an important factor: although self-attention is effective in most scenarios, the selection of the attention mechanism should be tailored to the specific task.

Additive attention mechanisms can capture context-specific interactions between modalities and demonstrate a high level of effectiveness in integrating various features by strategically emphasising pertinent components within a given sequence. The utilisation of additive attention for selective and context-aware feature fusion can offer significant advantages over the self-attention mechanism, especially when consolidating data from various modalities or feature sources.

Given these limitations, it is crucial to acknowledge that the effectiveness of the integrated feature and decision fusion methods should not be attributed solely to the underlying concept, but rather to the particular implementation and experimental decisions made. Considering the success of vision transformer-based feature fusion methods and the lower computational cost of the proposed methods, hybrid methods can be further investigated. Nevertheless, the available techniques for implementing hybrid approaches are currently limited, and the optimal integration of these approaches remains insufficiently investigated.

Another benefit of employing exclusively transformer-based neural networks is their rapid learning speed and absence of sequential operations, in contrast to ConvLSTM. Despite the promising results achieved by feature fusion-based video architectures, those architectures suffer severely from significant memory and computational costs.

The selection of an attention mechanism plays a crucial role in determining the ability of a model to effectively identify and highlight important information in the time-based sequences of video data.

V. CONCLUSION

This paper proposes novel approaches for feature and decision fusion at both single and multiple levels for video classification. The proposed multi-stream fusion architectures, combining spatial and temporal inputs, have demonstrated promising effectiveness across conducted experiments.

Feature-level fusion has outperformed decision-level fusion where all features through feature fusion resulted in the highest accuracy. Utilising additive attention within vision transformer based architectures has significantly improved video classification performance. The proposed multi-stream decision fusion method has achieved comparable results with reduced resource demands, emphasising its efficiency.

While the CNN network architecture employed in this study was not the state-of-the-art, the focus on fusion methods shows potential impact of the advanced designs on effectiveness. Future research should explore the scalability and resilience using larger datasets to ensure robustness.

In conclusion, the proposed fusion strategies, the vision transformer based approach in particular, have demonstrated great potential in effectively tackling the complex nature of video classification. Through the combination of feature fusion and decision fusion, our approach effectively enhances accuracy and offers practical implications for applications such as action identification, surveillance, among others. Further research endeavours will be focused on developing more powerful and practical deep learning architectures for video classification based on our findings in this study. Practical implementation of video classification methods in real-world circumstances remains a direction for future research.

REFERENCES

[1] C. D. D. Monteiro, C. M. Mathew, R. Gutierrez-Osuna, and F. Shipman, "Detecting and identifying sign languages through visual features," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2016, pp. 287–290.

[2] S. K. Shah and D. Shah, "Comparative study of image fusion techniques based on spatial and transform domain," *Int. J. Innov. Res. Sci., Eng. Technol.*, vol. 3, no. 6, pp. 10168–10175, 2014.

[3] B. Balachander and D. Dhanasekaran, *Comparative Study of Image Fusion Techniques in Spatial and Transform Domain*. Delhi, India: Asian Res. Publishing, 2016.

[4] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, Jul. 2018.

[5] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020.

[6] L. Wang, D. Q. Huynh, and M. R. Mansour, "Loss switching fusion with similarity search for video classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 974–978.

[7] T. Bi, D. Jarnikov, and J. Lukkien, "Video representation fusion network for multi-label movie genre classification," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9386–9391.

[8] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3560–3569.

[9] G. Xiao, D. P. Bavirisetti, G. Liu, and X. Zhang, "Decision-level image fusion," in *Image Fusion*. Cham, Switzerland: Springer, 2020, pp. 149–170.

[10] L. Wang, H. Zhou, S.-C. Low, and C. Leckie, "Action recognition via multi-feature fusion and Gaussian process classification," in *Proc. Workshop Appl. Comput. Vis. (WACV)*, Dec. 2009, pp. 1–6.

[11] C.-J. Lin, C.-H. Lin, and S.-Y. Jeng, "Using feature fusion and parameter optimization of dual-input convolutional neural network for face gender recognition," *Appl. Sci.*, vol. 10, no. 9, p. 3166, 2020.

[12] L. Chen, K. Bo, F. Lee, and Q. Chen, "Advanced feature fusion algorithm based on multiple convolutional neural network for scene recognition," *Comput. Model. Eng. Sci.*, vol. 122, no. 2, pp. 505–523, 2020.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[14] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, 2018.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[16] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–12.

[20] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[21] Z. Shao, Z. Hu, J. Yang, and Y. Li, "Multi-stream feature refinement network for human object interaction detection," *J. Vis. Commun. Image Represent.*, vol. 86, Jul. 2022, Art. no. 103529.

[22] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4305–4314.

[23] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "ActionVLAD: Learning spatio-temporal aggregation for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3165–3174.

[24] Z. Weng and Y. Guan, "Trajectory-aware three-stream CNN for video action recognition," *J. Electron. Imag.*, vol. 28, no. 2, p. 1, Dec. 2018, Art. no. 021004.

[25] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J. T. Zhou, and X. Bai, "Action recognition for depth video using multi-view dynamic images," *Inf. Sci.*, vol. 480, pp. 287–304, Apr. 2019.

[26] Z. Li, Z. Zheng, F. Lin, H. Leung, and Q. Li, "Action recognition from depth sequence using depth motion maps-based local ternary patterns and CNN," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 19587–19601, Jul. 2019.

[27] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 102–106.

[28] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 807–811, Mar. 2018.

[29] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 9532–9545, 2020.

[30] Q. Wu, Z. Wang, F. Deng, Z. Chi, and D. D. Feng, "Realistic human action recognition with multimodal feature selection and fusion," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 4, pp. 875–885, Jul. 2013.

[31] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[32] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[33] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *Int. J. Comput. Vis.*, vol. 130, no. 5, pp. 1366–1401, May 2022.

[34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2015, pp. 4489–4497.

[35] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[36] L. Jing, Y. Ye, X. Yang, and Y. Tian, "3D convolutional neural network with multi-model framework for action recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1837–1841.

[37] Y. Wang, Y. Xiao, F. Xiong, W. Jiang, Z. Cao, J. T. Zhou, and J. Yuan, "3DV: 3D dynamic voxel for action recognition in depth video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 508–517.

[38] Y. Wang, X. J. Shen, H. P. Chen, and J. X. Sun, "Action recognition in videos with spatio-temporal fusion 3D convolutional neural networks," *Pattern Recognit. Image Anal.*, vol. 31, no. 3, pp. 580–587, Jul. 2021.

[39] A. Ulhaq, N. Akhtar, G. Pogrebna, and A. Mian, "Vision transformers for action recognition: A survey," 2022, *arXiv:2209.05700*.

[40] E. Shabaninia, H. Nezamabadi-pour, and F. Shafizadegan, "Transformers in action recognition: A review on temporal modeling," 2022, *arXiv:2302.01921*.

[41] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in *Proc. 3rd Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds. San Diego, CA, USA, 2015.

[42] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," 2014, *arXiv:1412.4729*.

[43] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4507–4515.

[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[45] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[46] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[47] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*.

[48] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[49] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4694–4702.

[50] A. Stergiou and R. Poppe, "Spatio-temporal FAST 3D convolutions for human action recognition," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 183–190.

[51] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[52] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017.

[53] A. Sanchez-Caballero, D. Fuentes-Jimenez, and C. Losada-Gutiérrez, "Exploiting the ConvLSTM: Human action recognition using raw depth video-based recurrent neural networks," 2020, *arXiv:2006.07744*.

[54] X. Liu, Y. Li, and Q. Wang, "Multi-view hierarchical bidirectional recurrent neural network for depth video sequence based action recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 10, Oct. 2018, Art. no. 1850033.

[55] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[56] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[57] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[58] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8741–8750.

[59] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3202–3211.

[60] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang, C. Sun, and C. Schmid, "Multiview transformers for video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3333–3343.

[61] B. Li, P. Xiong, C. Han, and T. Guo, "Shrinking temporal attention in transformers for video action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1263–1271.

[62] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6836–6846.

[63] J. He and S. Gao, "TBSN: Sparse-transformer based Siamese network for few-shot action recognition," in *Proc. 2nd Inf. Commun. Technol. Conf. (ICTC)*, May 2021, pp. 47–53.

[64] O. Moutik, H. Sekkat, S. Tigani, A. Chehri, R. Saadane, T. A. Tchakoucht, and A. Paul, "Convolutional neural networks or vision transformers: Who will win the race for action recognitions in visual data?" *Sensors*, vol. 23, no. 2, p. 734, Jan. 2023.

[65] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[66] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.

[67] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[68] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, 2004, pp. 32–36.

[69] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *Center for Res. Comput. Vis.*, vol. 2, pp. 1–11, Jan. 2012.

[70] R. J. Freund, D. Mohr, and W. J. Wilson, *Statistical Methods*. New York, NY, USA: Academic, 2010.

[71] G. V. Glass, "Testing homogeneity of variances," *Amer. Educ. Res. J.*, vol. 3, no. 3, pp. 187–190, May 1966.

[72] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[73] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[74] S. N. Gowda, M. Rohrbach, and L. Sevilla-Lara, "SMART frame selection for action recognition," 2020, *arXiv:2012.10671*.

[75] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 20–36.

RUKIYE SAVRAN KIZILTEPE received the B.Sc. degree from Hacettepe University, Ankara, in 2014, and the M.Sc. and Ph.D. degrees from the School of Computer Science and Electronic Engineering, University of Essex, Colchester, in 2017 and 2022, respectively. She is currently an Assistant Professor with the Department of Software Engineering, Karadeniz Technical University. Her research interests include machine learning, video processing, computer vision, and video understanding using deep learning techniques.

JOHN Q. GAN received the B.Sc. degree in electronic engineering from Northwestern Polytechnic University, China, in 1982, and the M.Eng. degree in automatic control and the Ph.D. degree in biomedical electronics from Southeast University, China, in 1985 and 1991, respectively. He is currently a Professor in artificial intelligence with the University of Essex, U.K. He has coauthored a book and published over 200 research articles. His research interests include machine learning, artificial intelligence, signal and image processing, data and text mining, pattern recognition, brain–computer interfaces, and intelligent systems.

JUAN JOSÉ ESCOBAR received the Ph.D. degree in computer science from the University of Granada, Spain, in 2020. He is currently a permanent Lecturer with the Department of Software Engineering, University of Granada. His research interests include code optimization, energy-efficient parallel computing, and workload balancing strategies on heterogeneous and distributed systems, especially in issues related to evolutionary algorithms and multi-objective feature selection problems.