RESEARCH ARTICLE

Spatio-temporal Feature Soft Correlation Concatenation Aggregation Structure for Video Action Recognition Networks

Fafa Wang[1] **and Shenglun Yi**[2,*]

[1] Beijing iQIYI Technology Co., Ltd., China
[2] Department of Information Engineering, University of Padua, Italy

Abstract

The efficient extraction and fusion of video features to accurately identify complex and similar actions has consistently remained a significant research endeavor in the field of video action recognition. While adept at feature extraction, prevailing methodologies for video action recognition frequently exhibit suboptimal performance in the context of complex scenes and similar actions. This shortcoming arises primarily from their reliance on uni-dimensional feature extraction, thereby overlooking the interrelations among features and the significance of multi-dimensional fusion. To address this issue, this paper introduces an innovative framework predicated upon a soft correlation strategy aimed at augmenting the representational capacity of features by implementing multi-level, multi-dimensional feature aggregation and concatenating the temporal features produced by the network. Our end-to-end multi-feature encoding soft correlation concatenation aggregation layer, situated at the temporal feature output terminal of the Video Action Recognition network, proficiently aggregates and integrates the output temporal features. This approach culminates in producing a composite feature that cohesively unifies multi-dimensional information, markedly enhancing the network's competency in differentiating analogous video actions. Empirical findings demonstrate that the approach delineated in this paper bolsters the efficacy of video action recognition networks, achieving a more thorough depiction of images, and yielding superior accuracy and robustness.

Keywords: video action recognition, soft correlation, spatio-temporal feature extraction, concatenation aggregation structure, bidirectional LSTM.

1 Introduction

In recent years, video action recognition has emerged as a crucial area of research in computer vision, driven by its broad range of applications, from surveillance and security to human-computer interaction and content recommendation systems [1]. The ability to accurately identify and differentiate actions within video sequences is essential for developing intelligent systems that can interact with and interpret the dynamic visual world as humans do. However, despite

Submitted: 17 September 2024
Accepted: 18 October 2024
Published: 25 October 2024

Vol. 1, **No.** 1, 2024.
10.62762/TSCC.2024.212751

***Corresponding author:**
✉ Shenglun Yi
shenglun@dei.unipd.it

Citation
Wang, F., & Yi, S. (2024). Spatio-temporal Feature Soft Correlation Concatenation Aggregation Structure for Video Action Recognition Networks. *ICCK Transactions on Sensing, Communication, and Control*, 1(1), 60–71.

© 2024 ICCK (Institute of Central Computation and Knowledge)

significant advances, existing methods often need help with complex scenes and similar actions, primarily due to their reliance on single-dimensional feature extraction and insufficient consideration of feature interrelationships and multidimensional fusion [2].

The importance of this topic stems from the growing demand for more precise and efficient video action recognition systems that can handle the nuances and intricacies of real-world scenarios. Traditional feature aggregation algorithms, while performing adequately under controlled conditions, frequently fall short when faced with the variability and complexity inherent in real-life videos. This gap highlights the need for innovative approaches to capture and utilize multiple dimensions of feature information, enabling a more comprehensive and discriminative representation of actions [3].

Despite the extensive research done in this field, there remains a notable gap in effectively aggregating and fusing features at multiple levels to enhance action recognition accuracy. Many existing studies have focused on improving individual aspects, such as spatial or temporal feature extraction, but fewer have explored holistic approaches that integrate multiple feature dimensions. For example, Wang et al. [4] presented a temporal segment network (TSN) that segments video into uniform parts and extracts temporal features for each segment individually, enhancing the temporal dimension's representation but largely overlooking the potential benefits of integrating spatial features. Similarly, Yang et al. [5] introduced a two-stream inflated 3D ConvNet (I3D), which provides a robust framework for capturing spatiotemporal features separately but does not address the holistic integration of these features.

Moreover, researchers like Tran et al. [6] have explored using 3D convolutional networks for learning spatiotemporal features. Their C3D framework significantly improves action recognition by simultaneously capturing spatial and temporal dynamics; however, it lacks a comprehensive method for multi-dimensional feature fusion. Carreira et al. [7] extended this concept with the I3D model that inflates 2D ConvNets, pre-trained on ImageNet, to 3D. This method achieved state-of-the-art results by focusing on volumetric spatiotemporal feature extraction, yet it still treated spatial and temporal features relatively independently. Another approach by Feichtenhofer et al. [8] proposed a SlowFast network architecture that processes video inputs at different frame rates, effectively balancing spatial and temporal information. This method moves closer to a holistic feature aggregation but does not fully exploit the interrelationships between different feature dimensions.

These studies collectively highlight advancements in specific aspects of feature extraction but also underscore the deficiencies in strategies for comprehensive feature aggregation and multi-dimensional fusion. By addressing this gap, our research aims to develop a method that captures and effectively integrates multi-dimensional features, thereby enhancing the discriminative power and accuracy of video action recognition networks. Therefore, our research aims to address the following questions: 1) In what manner can multi-dimensional features be aggregated and fused to enhance the discriminative capability of video action recognition networks? 2) What benefits does a soft-association strategy offer concerning feature aggregation and fusion? 3) Is it possible for an end-to-end multi-feature encoding layer to improve the network's proficiency in distinguishing between similar actions within video sequences?

This paper introduces an innovative end-to-end multi-feature encoding soft-association concatenation aggregation layer. Our contributions are as follows:

1. We propose a novel feature aggregation method that utilizes soft-association strategies to enhance the interrelationships between different feature dimensions.

2. Our method integrates seamlessly with existing video action recognition networks, enhancing their capability to effectively process and aggregate temporal features.

3. We demonstrate that the generated composite feature can represent the feature distribution from multiple dimensions, providing a more comprehensive image description.

We show that our experimental results validate the effectiveness of the proposed method, showcasing its potential to advance the field of video action recognition and provide a robust solution for handling complex and nuanced video data. The organization of this paper is as follows: Section 2 offers a comprehensive review of the existing literature in the field. Section 3 elaborates on the research methodology, encompassing the preprocessing for data augmentation, the extraction of spatiotemporal features from videos using the LI3D network, and the utilization of the LI3D-BiLSTM architecture for spatiotemporal feature extraction. Section 4 outlines the experimental setup, including the dataset employed and the presentation of the findings and analysis. Conclusively, Section 5 summarizes the paper, highlighting the principal discoveries and the implications of the research for the field.

2 Related Work

As the complexity and volume of video data continue to increase, video scenarios have evolved from simple, isolated settings to more diverse and intricate environments, bringing video-based motion pattern recognition closer to real-world applications. Consequently, the primary challenge in this field has shifted to extracting and identifying motion features in noisy environments. Modern algorithms focus on capturing motion characteristics and strive to improve accuracy under computational efficiency constraints.

In earlier studies, Saha et al. [9] combined Convolutional Neural Networks (CNN) with Support Vector Machines (SVM) to classify human motion states, highlighting the potential of integrating deep learning with traditional machine learning techniques. Building on this, Karpathy et al. [10] used a fixed-window approach to process stacked image features in video frames, further enhancing CNN's action recognition capabilities. Funke et al. [11] applied 3D CNNs to the Sports-1M dataset, successfully capturing spatiotemporal features in videos, thus significantly improving recognition accuracy. Additionally, Yao et al. [12] conducted a comprehensive survey of CNN-based approaches for action recognition, summarizing key advancements and emphasizing strategies that enhance spatiotemporal feature extraction.

In recent years, the combination of CNN and Long-Short-Term Memory (LSTM) networks has gained significant attention for motion pattern recognition. CNNs excel at extracting features from individual frames, while LSTMs are adept at handling the temporal relationships between consecutive frames. Donahue et al. [13] proposed an LSTM-based autoencoder model that showed a marked improvement in classification accuracy. Zou et al. [14] applied regularization techniques to neural networks, optimizing feature-sharing mechanisms to ensure better collaboration among different features.

The introduction of Transformer architectures has opened new possibilities for video understanding. Bertasius et al. [15] introduced the TimeSformer model, showcasing the advantages of self-attention mechanisms in capturing long-range spatiotemporal dependencies. However, due to its high computational complexity, this model faced challenges in practical applications. To address these issues, Wu et al. [16] introduced the Memory-augmented Multiscale Vision Transformer (MemViT), which enhances long-term video recognition by integrating memory mechanisms with hierarchical pooling to achieve high efficiency and strong performance.

The rise of self-supervised learning has provided new directions for utilizing large-scale unlabeled video data. Tong et al. [17] introduced the VideoMAE model, which notably enhanced data efficiency and inspired future research in unsupervised learning. At the same time, Yang et al. [18] proposed a temporal shift attention mechanism that effectively combines the advantages of self-attention and the Temporal Shift Module (TSM), enabling the capture of long-term dependencies at a low computational cost.

The current trends suggest that future research will focus on efficiently extracting and combining spatiotemporal features in video data. This will involve utilizing Transformer architectures and self-supervised learning techniques to improve the performance and effectiveness of video understanding.

This study presents an advanced methodology for the recognition of video actions, grounded in a lightweight Inception-3D networks (LI3D) architecture aimed at the extraction of spatiotemporal features. Moreover, it introduces a soft-association feature aggregation module designed to augment the recognition accuracy of critical video actions. In order to further augment the network's capability in recognizing temporal features within video data, this investigation expands upon the foundational work of Bidirectional Long Short-Term Memory networks (Bi-LSTM). The sequences of video features obtained by the LI3D network are subjected to processing through Bi-LSTM for contextual association, thereby bolstering the network's proficiency in representing temporal data features.

In addition, this paper introduces a novel structure based on a soft-association strategy. This approach aims to enhance the representational capacity of features by performing multi-level and multi-dimensional feature aggregation and concatenation on the temporal features output by the network. Our proposed method leverages the concept of soft association, allowing for flexible and adaptive connections between different feature dimensions. Unlike hard association methods that enforce rigid relationships, our soft-association strategy enables the network to learn and adjust the strength of connections between various feature aspects dynamically. This adaptability is crucial when dealing with the complex and often ambiguous nature of actions in real-world video sequences.

The multi-level aspect of our approach involves processing features at different scales of abstraction. By considering both low-level details and high-level semantic information, we ensure a comprehensive representation of the video content. This multi-level processing helps in capturing both fine-grained movements and overarching action patterns. Furthermore, our method emphasizes multi-dimensional feature aggregation, recognizing that video actions are characterized by various attributes such as spatial configuration, temporal dynamics, and contextual information. We create a richer and more informative feature representation by explicitly designing our structure to aggregate and fuse these multiple dimensions.

The concatenation step in our approach effectively combines these multi-level and multi-dimensional features. Rather than simply averaging or pooling features, concatenation preserves the distinct information from each level and dimension, allowing the subsequent layers of the network to leverage this comprehensive feature set. Ultimately, this sophisticated feature aggregation and concatenation process significantly enhances the network's ability to distinguish between similar actions and handle complex scenarios. By improving the overall representational power of the features, our method aims to push the boundaries of accuracy and robustness in video action recognition tasks.

3 Methodology

3.1 The Traditional Feature Encoding

Image characteristics can generally be categorized into global and local features. Global features encapsulate the comprehensive content information of an image, such as color, texture, and shape, whereas local features encompass the specific information related to the minute details of an image, including corners, edges, and lines. In contrast to global features, local features are characterized by their abundance, low correlation among features, and reduced susceptibility to occlusions. In recent years, local features have gained widespread application in domains such as face recognition, 3D reconstruction, object recognition, object detection, and tracking. In the realm of motion pattern recognition algorithms centered on video content, the re-encoding of local features to enhance the representation of spatial and temporal relationships within each frame directly influences the efficacy of recognition and classification algorithms.

3.1.1 Bag of Feature

Bag of Features (BOF) is a feature aggregation strategy derived from the Bag of Words model. Similar to the Bag of Words model, the idea behind the BOF algorithm is to create a dictionary that contains all the local features of an image. These features are obtained through a clustering algorithm to determine many cluster centers. These cluster centers are usually highly representative. For example, for faces, although the features such as eyes and noses of different people are not the same, they often have commonalities, and these cluster centers represent such commonalities. The aggregation of features at these cluster centers can achieve a higher-dimensional expression relative to local features. We combine these cluster centers to form a dictionary. For each SIFT feature in the image, we can find the most similar cluster center in the dictionary and count the occurrences of these cluster centers to obtain a histogram of the distribution of feature cluster centers. For images of different categories, there will be a significant difference in the distribution of this histogram. Therefore, based on these differently distributed feature word bands, some classification models can be trained, and they can be used to classify images.

The principle of the BOF algorithm is actually very simple, which can be summarized in one sentence: create a dictionary, generate a vector, and finally count the frequency of the words. The specific algorithm for constructing BOF is shown in Algorithm 1.

Utilizing the Bag of Features (BOF) feature aggregation strategy, we can obtain the frequency distribution of local features in a higher dimension relative to the cluster centers. This is beneficial for feature recognition and classification. However, the BOF algorithm has obvious drawbacks:

1. It does not take into account the positional relationships between features at all, while positional information is very important for understanding images.

Algorithm 1: BOF Feature Aggregation Algorithm Process

```
Input: SIFT descriptors extracted from the training data ($M$ images)
Output: Word frequency vector after aggregating local features through the BOF algorithm
Objective: To enhance the expression ability of local features and abstract features that can comprehensively represent image content.
  1. Preprocess the $M$ images, including image cropping, scaling, and other preprocessing procedures.
  2. Extract SIFT features using the SIFT corner detection algorithm in OpenCV for each image. Each image can extract several SIFT descriptors with a dimension of 128. Assume that a total of $N$ 128-dimensional SIFT features are extracted from the $M$ images.
  3. Perform K-means clustering on the $N$ SIFT features extracted in step 2, grouping the $N$ SIFT keypoints into $K$ cluster centers, so that the similarity within the same cluster center is high, while the similarity between different cluster centers is low.
  4. Calculate the distance of all SIFT keypoints in each image to the $K$ cluster centers, map all SIFT keypoints to the nearest cluster center, and increment the count of that cluster center by 1.
```

Algorithm 2: Fisher Vector (FV) Feature Aggregation Process

```
Input: SIFT descriptors extracted from the training data ($M$ images)
Output: Feature vector obtained after describing local features with the FV algorithm
Objective: To enhance the expressive power of local features and to abstract a feature vector that can comprehensively represent the content of images.
  1. Extract $T$ descriptors from an image, each descriptor being $D$-dimensional, to obtain the image descriptors $X = \{x_1, \cdots, x_T\}$.
  2. Using a linear combination of $K$ Gaussian Mixture Models (GMM) to approximate the distribution of $T$ descriptors $p(x_t \mid \lambda) = \sum_{i=1}^{K} w_i p_i(x_t \mid \lambda_i)$, where $\lambda$ represents the prior values obtained in advance through GMM.
  3. The Fisher Vector for the image is obtained by taking the partial derivative of the Gaussian model, that is $\{f_{w_i}^{-1/2}\partial L(X \mid \lambda)/\partial w_i, f_{\mu_i^d}^{-1/2}\partial L(X \mid \lambda)/\partial \mu_i^d, f_{\sigma_i^d}^{-1/2}\partial L(X \mid \lambda)/\partial \sigma_i^d\}$.
  4. Perform steps 2 and 3 on all images in the training set to obtain a set of FV-represented features for the training set.
```

2. BOF only focuses on the quantity of key features, which is a zero-order statistic. The resulting feature vector is sparse, so the feature expression is not rich enough.

3.1.2 *Fisher Vector*

Fisher Vector (FV) effectively addresses the sparsity issue of the BOF feature vector, significantly increasing the dimensionality of image features. In the FV feature representation, in addition to the zero-order features, it also includes first-order (mean) and second-order (variance) features. Therefore, FV can more fully represent the content of images. Essentially, FV expresses an image using the gradient vector of the likelihood function. Since actual data distributions often follow a mixture of Gaussian distributions, for an image with N descriptors, it is assumed that these features conform to a certain distribution and that these distributions are independent of each other. We can use a linear combination of N Gaussian distributions to approximate the distribution of these features, that is, to represent the probability distribution of a sample (an image) as the product of the probability distributions of each feature dimension. This feature aggregation strategy can better describe the actual situation of the features. Moreover, taking the logarithm of the product of probability distributions turns it into a summation form, which greatly reduces computational complexity. The specific steps of the FV algorithm are shown in Algorithm 2.

3.1.3 *VLAD*

VLAD, similar to BOF, only considers the closest cluster center to the feature and records the distance between the feature and its nearest cluster center. At the same time, like FV, VLAD takes into account the information of each dimension of the feature point, providing a more microscopic depiction of the image's local features. VLAD can aggregate statistical data of local descriptors on the image and store the residuals and sums of local descriptors to their corresponding cluster centers. The specific process of the VLAD algorithm is shown in Algorithm 3. VLAD ensures a detailed depiction of data features while

```
Algorithm 3: VLAD Feature Aggregation Process
Input: SIFT descriptors extracted from the training data ($M$ images)
Output: Feature vector obtained after describing local features with the VLAD algorithm
Objective: To compensate for the information loss in BOF and FV, and to abstract a feature vector
    that can comprehensively represent the content of images.
  1. Local feature descriptors are extracted using traditional methods or deep learning methods,
     represented by $x$.
  2. Use the k-means clustering algorithm to obtain a codebook with $k$ cluster centers $C = \{c_1, \cdots, c_k\}$.
  3. Assign each descriptor to the closest codebook entry, and after the assignment, the feature space
     will be divided into multiple Voronoi regions.
  4. Calculate the sum of residuals of features to their respective cluster centers within each Voronoi
     region $x - c_i$,
         $v_{ij} = \sum_{l=1}^{N} a_i(x_l)(x_l(j) - c_i(j)),$
     where $i$ represents the number of cluster centers, $j$ represents the dimensions of each cluster center
     and local feature descriptor, $l$ represents the total number of local feature descriptors, and $a_i(x_l)$
     represents the probability that the $l$-th local feature descriptor is associated with the $i$-th
     cluster center.
  5. Finally, normalize the residual vector to obtain the final expression form of VLAD $v = \frac{v}{\|\mathbf{v}\|_2}$.
```

effectively reducing the computational load of feature aggregation strategies. This feature aggregation strategy balances the richness of feature expression and the improvement of computational performance, with VLAD expressing the relationship between feature points and cluster centers in the form of residuals.

However, these traditional methods all share a common drawback: the cluster centers obtained using algorithms similar to k-means cannot be optimized according to the needs of the data, which is the limitation of traditional approaches. Therefore, based on the study of traditional feature aggregation strategies, this paper improves the traditional feature aggregation algorithms in light of the idea that deep learning can optimize parameters through back propagation according to the data distribution.

4 Soft Assignment

The traditional feature aggregation strategies mentioned in the previous section all utilize the K-means algorithm to obtain the codebook for feature aggregation, hence the cluster centers are located at the center of each aggregation area, which clearly cannot ensure that the distance from local feature descriptors to the cluster centers is minimized. To address this issue, this paper proposes a soft association feature aggregation strategy, the main approach of which is to rewrite the affiliation degree a of features to cluster centers in traditional aggregation strategies into a soft association form. In the BoF, Fisher Vector, and VLAD strategies, the affiliation degree of features to cluster centers is a constant a. When $a = 1$, it indicates that the current feature belongs to that cluster center; when $a = 0$, it indicates that the current feature does not belong to that cluster center. The soft association strategy proposed in this paper represents the affiliation degree with $a_i(x_l)$, as shown in Equation (1).

$$\bar{a}_i(x_l) = \frac{e^{-\alpha\|x_l - c_i\|^2}}{\sum_{i'} e^{-\alpha\|x_l - c_{i'}\|^2}} \tag{1}$$

where α is a constant between 0 and 1, representing the probability value of the possibility that the local feature descriptor x_l belongs to the cluster center c_i. The value of α is inversely proportional to the distance from the local descriptor to the cluster center. Simplifying Equation (1) yields Equation (2).

$$\bar{a}_i(x_l) = \frac{e^{w_i^T x_l + b_i}}{\sum_{i'} e^{w_{i'}^T x_l + b_{i'}}} \tag{2}$$

During the training process, we first initialize the cluster centers, which vary slightly depending on the traditional clustering methods. For the three clustering methods of BoF, FV, and VLAD, we initialize the cluster centers as the frequency distribution of features, the mean and variance distribution of features, and the distribution of the sum of residuals of features to each element of the cluster centers, respectively. During the network training, the affiliation degree of features to a certain cluster center is represented by a probability value between 0 and 1, allowing the affiliation degree between features and cluster centers to be adjusted through the backpropagation and gradient descent of the neural network. That is, these initialized cluster centers will be adjusted according to the differences in each batch of input data, making the cluster centers fit the data distribution better, thereby making the aggregation degree of the same category features more obvious and gradually increasing the distance between different category features.

We add the affiliation degrees processed by soft association to the local aggregation formula, and the complete expression after improvement is shown in Equation (3). The improved algorithms are called soft-BOF, soft-FV, and soft-VLAD.

$$v_{i,j} = \sum_{l=1}^{N} \bar{a}_i(x_l)\left(x_l(j) - c_i(j)\right) \tag{3}$$

where w_i, b_i, c_i represent the weight, bias, and vector representation of the initialized cluster centers, respectively. The parameters w_i, b_i, c_i can be continuously updated during the training phase, which makes the feature aggregation layer modified by the soft association strategy more flexible, enabling the cluster centers to move in a direction that further reduces the distance between them and the features.

4.1 Multi-Feature Encoding by Soft Concatenation Aggregation

Soft-BoF implements the 0th-order frequency statistics of feature distribution, soft-FV implements the 1st-order mean and 2nd-order variance statistics of feature distribution, and soft-VLAD directly implements the residual statistics of feature distribution with respect to the cluster centers. After studying the above three feature aggregation methods, we constructed an end-to-end multi-feature encoding soft association concatenation aggregation algorithm. This method uses the three statistical quantities obtained from soft-BoF, soft-FV, and soft-VLAD as cluster centers, and uses the soft association strategy to obtain three forms of feature representation. Then, these three types of features are concatenated together to serve as a new feature descriptor. This soft association concatenation aggregation layer can continuously adjust the cluster centers using the network's backpropagation, making the cluster centers closer to the distribution of the actual data. That is, features belonging to the same category are more compactly aggregated, and features not belonging to the same category are more dispersed. This will be more conducive to recognizing similar actions in videos and improving the network's recognition and classification performance.

The multi-feature encoding soft association concatenation algorithm can fuse features with three levels of aggregation, finally obtaining a fused feature that can represent the feature distribution from multiple dimensions. This feature can provide a more comprehensive description of the image from multiple dimensions, and the specific fusion process is as follows.

1. Fuse the features of the three soft association clustering methods, where the features are aggregated through the three clustering methods. The output dimension is $batch\_size * output\_dim$, where $batch\_size$ is the total number of input samples per batch, and $output\_dim$ is the output feature dimension after clustering. After the fusion operation, the feature dimension becomes $(merge\_classes * batch\_size) * output\_dim$, where $merge\_classes$ is the number of feature aggregation methods.
2. Perform a reshape operation on the fused features, changing the feature dimension from the original $(merge\_classes * batch\_size) * output\_dim$ to $batch\_size * merge\_classes * output\_dim$.
3. Perform an averaging operation on the feature dimension $merge\_classes$, reducing the dimension of $merge\_classes$ to 1D.
4. Finally, use the squeeze module in TensorFlow to compress the 1-dimensional part of the feature dimension, at which point the final fused feature dimension becomes $batch\_size * output\_dim$ again, the output dimension remains unchanged, and the features include the three types of aggregation representations.

The final expression after fusing the three feature aggregation methods is shown in Equation (4), where i represents the number of cluster centers, j represents the dimension of features and cluster centers, l represents the number of features per batch, $classes$ represents the number of feature aggregation methods, and the aggregated features are represented as V.

$$V(j,i) = \sum_{classes=1}^{M} \sum_{l=1}^{N} \frac{e^{w_i^T x_l + b_i}}{\sum_{i'} e^{w_{i'}^T x_l + b_{i'}}} \left(x_l(j) - c_i(j)\right) \tag{4}$$

Figure 1 describes the specific process of the multi-feature encoding soft association concatenation algorithm, with the red box section comparing the cluster centers of this algorithm with those of traditional feature aggregation algorithms. In it, the black star represents the original cluster center, and the red star represents the position of the cluster center after adjustment by the network's back propagation during the training process in the multi-feature encoding soft association concatenation algorithm. It is evident that the adjusted cluster center is closer to the local features than the original cluster center, resulting in a stronger feature aggregation representation capability.

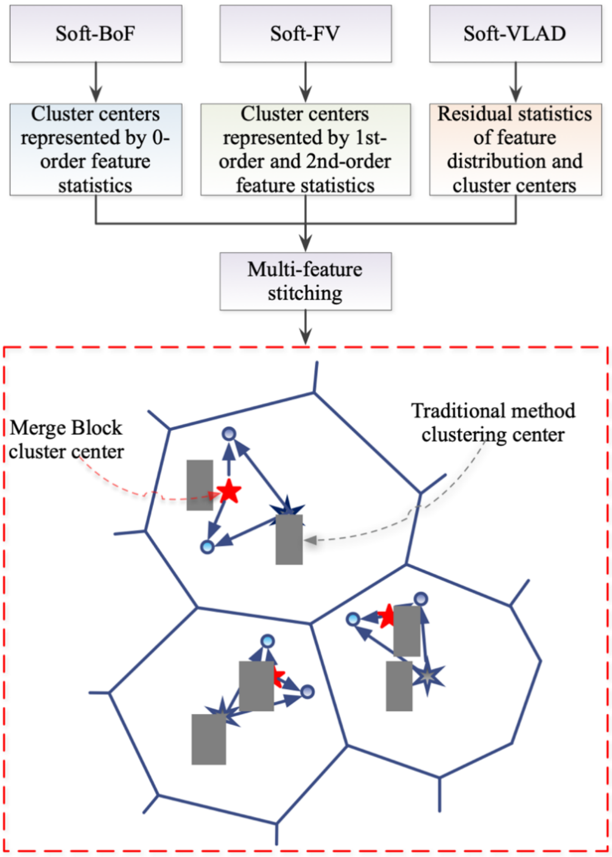

Soft-BoF
Soft-FV
Soft-VLAD
Cluster centers represented by 0-order feature statistics
Cluster centers represented by 1st-order and 2nd-order feature statistics
Residual statistics of feature distribution and cluster centers
Multi-feature stitching
Merge Block cluster center
Traditional method clustering center

Figure 1. Merge block cluster centers and traditional cluster centers.

5 Experiments and Analysis

The separability of features is a prerequisite for the effectiveness of an algorithm. If the extracted features are inseparable, then it is meaningless to blindly carry out network training. Many factors can cause data to be inseparable. In the field of video content understanding, the main factors are: 1) The data itself does not have separability, or the features between the data are too similar, resulting in unclear distinguishability; 2) The data is separable, but there is a problem with the feature extraction algorithm, which prevents the extracted features from adequately reflecting the distribution of the original data. This situation can also lead to inseparable features. It can be seen that the study of feature separability and the visualization of features are crucial for adjusting the algorithm structure and deeply analyzing the data. Dimensionality reducing the extracted features to two or three dimensions and visualizing them helps to analyze which data have clear separability and which data have unclear distinguishability. We can then make targeted algorithm improvements for data with less obvious separability, which is crucial for reasonably optimizing the algorithm.

Figure 2 (left) shows the feature distribution of the 101 categories in the UCF101 dataset extracted from the I3D-BiLSTM model without multi-feature encoding soft association. Figure 2(right) shows the feature distribution of the same dataset extracted from the I3D-BiLSTM model after multi-feature encoding soft association. It is not difficult to see that after the

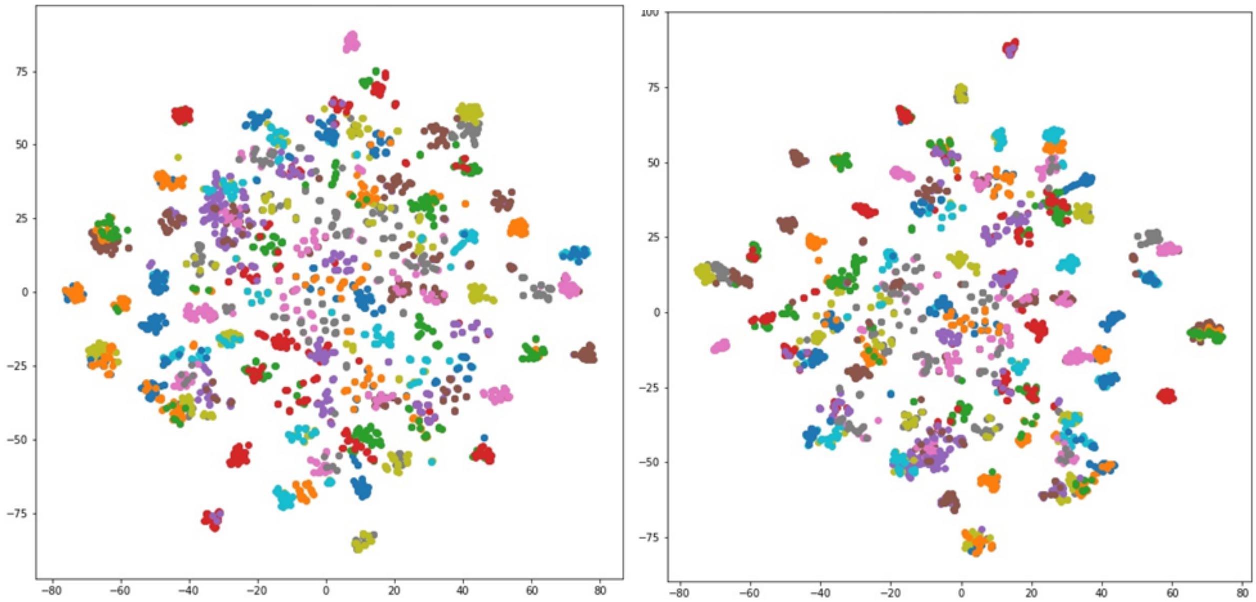


Figure 2. Feature distribution before and after multi-feature encoding soft association concatenation.

multi-feature encoding soft association strategy, the distance between different category features is further apart, and the aggregation of the same category features is more compact. This indicates that the multi-feature encoding soft association strategy can effectively improve the network's ability to recognize video features.

Figure 3 shows the visualization of feature distribution for some similar videos in the UCF101 dataset. The left side is the feature distribution extracted by I3D-BiLSTM, and the right side is the feature distribution after multi-feature encoding soft association strategy is applied to the features extracted by I3D-BiLSTM. Each color in the graph represents a different category of video. It is not difficult to see from Figure 3 that before feature aggregation, the distribution of features is quite scattered, the aggregation of the same category features is not compact enough, and the distance between different categories' features is also close. Overall, although it is somewhat separable to a certain extent, some categories are still not separable. The red box in the graph represents two categories in UCF101, Boxing PunchingBag and Boxing SpeedBag. These two categories are inherently similar in appearance, so without feature aggregation, it is difficult to distinguish between these two categories. The feature distribution after feature aggregation is shown on the right side of Figure 3. It can be seen that after the feature aggregation strategy, the aggregation of each category is more obvious, and the features of different categories are easier to distinguish. In particular, the previously inseparable Boxing PunchingBag and Boxing SpeedBag categories have significantly improved separability after the feature aggregation strategy.

Figure 4 shows the accuracy curves of the network training process before and after feature aggregation with 50 iterations. The green curve represents the accuracy curve of the proposed model LI3D-BiLSTM without implementing the multi-feature encoding soft association strategy, referred to as LI3D-BiLSTM. The red curve represents the accuracy curve of LI3D-BiLSTM after applying the multi-feature encoding soft association strategy, indicated as LI3D-BiLSTM(*). It can be observed that after applying the multi-feature encoding soft association strategy, the model's accuracy has significantly improved compared to before. Moreover, the initial accuracy of the model is already much higher than that of the basic model, and around the 20th iteration, the proposed model tends to stabilize.

Figure 5 compares the loss functions of the I3D-BiLSTM model output during the training process with and without the multi-feature encoding soft association strategy under 50 iterations. The red curve represents the loss function curve of the network after applying the feature aggregation strategy, while the green curve represents the loss function curve without the feature aggregation strategy. It can be observed that after applying the feature aggregation strategy, the rate of decrease in the loss function is significantly faster than that of the baseline model. This indicates that after feature aggregation, the network model can converge more quickly, resulting in a more stable model with better robustness.

Table 1 shows the test set accuracy, precision, and recall for the scenarios with and without feature aggregation over 50 iterations. The metrics in Table 1 clearly show that after applying feature aggregation, the network model demonstrates better recognition performance on the test set. There is an improvement of approximately three percentage points across the three metrics: accuracy, precision, and recall.

Table 1. Comparison of Test Set metrics before and after feature aggregation.

Index	Before	After
Accuracy	0.90	0.92
Precision	0.91	0.93
Recall	0.90	0.91

Table 2 compares the number of parameters, testing time, and testing accuracy of the I3D-BiLSTM network after incorporating the multi-feature encoding soft association feature aggregation module under different clustering center parameters. It can be seen from Table 2 that when the clustering center is set to 1024 dimensions, the number of parameters in the network becomes excessively large, and the testing time is also relatively long. This poses a significant challenge to the usability of the network. As the number of clustering centers decreases, both the number of parameters and the testing time of the network model decrease substantially. However, when the clustering center is reduced to 64 dimensions, it leads to a decrease in testing accuracy, which does not meet the performance requirements for network recognition. When the number of clustering centers in the feature aggregation module is reduced to 128 dimensions, the number of parameters is reduced

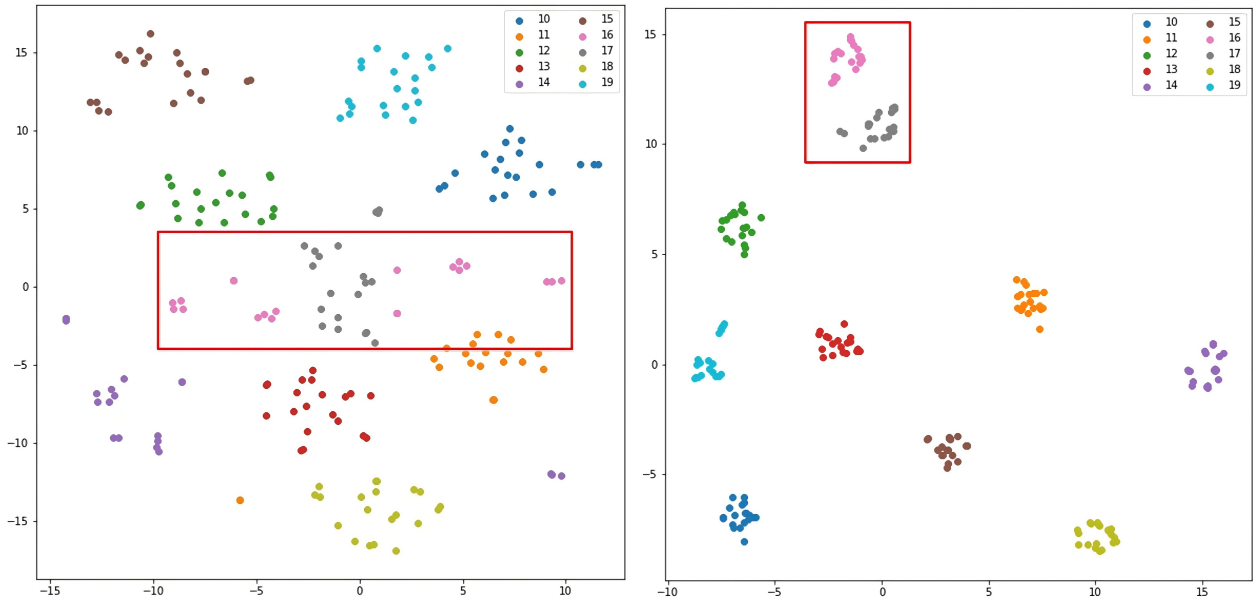


Figure 3. UCF101 partial category feature clustering visualization.

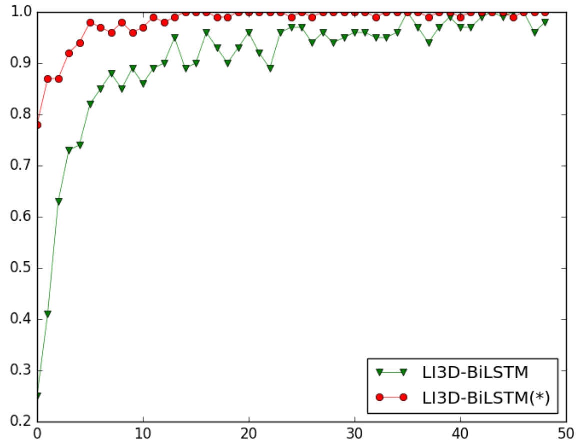


Figure 4. Accuracy curves of the training process before and after feature aggregation strategy.

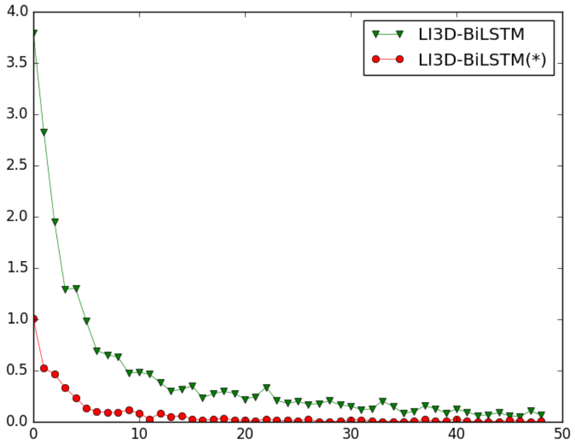


Figure 5. Loss function curves before and after feature aggregation strategy.

by 92.47% and 84.21% compared to 1024 and 256 dimensions, respectively, and the testing time gets closer to real-time requirements. At this point, the network structure can maintain relatively high testing accuracy while ensuring the minimum number of parameters. Thus, having 128 clustering centers in the feature aggregation module is considered the optimal strategy. In subsequent experiments, the default setting for the number of clustering centers in the multi-feature encoding soft association feature aggregation module will be 128 dimensions, which allows for the least aggregate time and higher aggregation efficiency.

6 Conclusion

In this paper, we conducted an in-depth study comparing traditional local feature encoding methods with modern deep learning-based encoding techniques. Through detailed analyses, we elucidated the strengths and weaknesses of both approaches. Building upon these insights and traditional feature aggregation algorithms, we introduced an innovative feature aggregation pooling layer utilizing a soft association strategy. This led to the development of a spatiotemporal feature soft-association concatenation aggregation module. Our module integrates three soft-association feature aggregation strategies: soft-Bag of Features (soft-BOF), soft-Fisher Vector

Table 2. Effect of the number of clustering centers on soft association splicing aggregation.

Cluster center	Parameters	Test time	Test accuracy
1024	156.78M	8.63s	0.918
256	74.71M	5.17s	0.921
128	11.80M	1.58s	0.921
64	9.64M	1.57s	0.899

(soft-FV), and soft-Vector of Locally Aggregated Descriptors (soft-VLAD). By doing so, we effectively combined the frequency distribution, mean-variance distribution, and residual distribution of features. This multi-faceted integration approach is optimized through the backpropagation of gradients within the network, allowing the clustering centers to be iteratively refined. Consequently, this enhances the clustering of features and facilitates better differentiation between feature categories.

Significantly, our module's capacity to unify and optimize different distribution types of features marks a considerable scientific advancement. We demonstrated, through comparative experiments, that selecting a dimensionality of 128 for the cluster center parameter of our module strikes an optimal balance. Under this configuration, our network model maintains high recognition accuracy while substantially reducing the number of network parameters and testing time. Additionally, our approach helps to mitigate the risks of overfitting, contributing to more robust and generalizable models. In conclusion, our proposed spatiotemporal feature soft-association concatenation aggregation module stands as a robust contribution to the field of video action recognition, offering enhanced feature encoding, improved differentiation, and practical efficiency. This work lays the groundwork for future explorations into more dynamic and adaptable feature aggregation strategies in complex video analysis tasks.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

Fafa Wang is an employee of Beijing iQIYI Technology Co., Ltd., China.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Kong, J., Wang, H., Wang, X., Jin, X., Fang, X., & Lin, S. (2021). Multi-stream hybrid architecture based on cross-level fusion strategy for fine-grained crop species recognition in precision agriculture. *Computers and Electronics in Agriculture, 185*, 106134. [CrossRef]

[2] Liang, Z., Zhou, R., Zhang, L., Li, L., Huang, G., Zhang, Z., & Ishii, S. (2021). EEGFuseNet: Hybrid unsupervised deep feature characterization and fusion for high-dimensional EEG with an application to emotion recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29*, 1913-1925. [CrossRef]

[3] Georgiou, T., Liu, Y., Chen, W., & Lew, M. (2020). A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision. *International Journal of Multimedia Information Retrieval, 9*, 135-170. [CrossRef]

[4] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20-36). Springer, Cham. [CrossRef]

[5] Yang, Z., An, G., Zhang, R., Zheng, Z., & Ruan, Q. (2023). SRI3D: Two-stream inflated 3D ConvNet based on sparse regularization for action recognition. *IET Image Processing, 17*(5), 1438-1448. [CrossRef]

[6] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489-4497).

[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] 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).

[9] Saha, A., Mazumdar, M., & Ghosh, A. (2019). Human motion recognition using CNN and SVM. *Journal of Ambient Intelligence and Humanized Computing, 10*(4), 1561574.

[10] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3128-3137).

[11] Funke, I., Bodenstedt, S., Oehme, F., von Bechtolsheim, F., Weitz, J., & Speidel, S. (2019, October). Using 3D convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition

in video. In *International conference on medical image computing and computer-assisted intervention* (pp. 467-475). Cham: Springer International Publishing.

[12] Yao, G., Lei, T., & Zhong, J. (2019). A review of convolutional-neural-network-based action recognition. *Pattern Recognition Letters, 118*, 14-22. [CrossRef]

[13] 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).

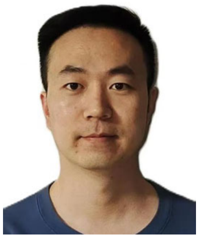
[14] Zou, J., Wang, D., & Li, X. (2020). Adaptive Regularization for CNNs. *Neural Networks, 134*, 151-159.

[15] Bertasius, G., Wang, H., & Torresani, L. (2021, July). Is space-time attention all you need for video understanding?. In *ICML* (Vol. 2, No. 3, p. 4).

[16] Wu, C. Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., & Feichtenhofer, C. (2022). Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 13587-13597).

[17] Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems, 35*, 10078-10093.

[18] Yang, J., & Yu, H. (2022). Temporal Shift Attention for Action Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 121-130.



Fafa Wang, a 2019 graduate of Beijing Technology and Business University with a degree in Control Theory and Control Engineering, currently serves as an Algorithm Engineer at iQIYI Technology Co., Ltd. in Beijing . Specializing in computer vision and large-scale model applications, Wang's work encompasses diverse areas, including image and video recognition, retrieval, and tracking; NLP-based danmaku comment recognition; large model-based empty shot material recognition; speaker recognition; and video orientation conversion. His background in control engineering, combined with his focus on cutting-edge AI technologies, enables Wang to contribute significantly to various projects at iQIYI, effectively bridging theoretical knowledge with practical applications in artificial intelligence and multimedia processing. (Email: wangfafa@qiyi.com)

Shenglun Yi is currently an Assistant Professor in the Department of Information Engineering at the University of Padova, Italy. He received his B.Eng. degree in Automation from Chongqing University, China 2016, followed by an M.Sc.Eng. degree in Control Engineering from Beijing Technology and Business University, China, in 2018. In 2022, he completed his Ph.D. in Control Science and Engineering at Beijing Institute of Technology, China. Dr. Yi's research interests encompass a range of topics, including robust estimation, information fusion, signal processing, and identification theory. His diverse educational background and current position at a prestigious Italian university demonstrate his international experience and expertise in the field of information engineering and control systems. (Email: shenglun@dei.unipd.it)