

Research Article

Channel Attention-Based Approach with Autoencoder Network for Human Action Recognition in Low-Resolution Frames

Elaheh Dastbaravardeh,¹ Somayeh Askarpour,² Maryam Saberi Anari ,² and Khosro Rezaee ³

¹Department of Control Engineering, Islamic Azad University of Mashhad, Mashhad, Iran

²Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran

³Department of Biomedical Engineering, Meybod University, Meybod, Iran

Correspondence should be addressed to Khosro Rezaee; kh.rezaee@meybod.ac.ir

Received 22 January 2023; Revised 27 October 2023; Accepted 12 December 2023; Published 4 January 2024

Academic Editor: Alexander Hošovský

Copyright © 2024 Elaheh Dastbaravardeh et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Action recognition (AR) has many applications, including surveillance, health/disabilities care, man-machine interactions, video-content-based monitoring, and activity recognition. Because human action videos contain a large number of frames, implemented models must minimize computation by reducing the number, size, and resolution of frames. We propose an improved method for detecting human actions in low-size and low-resolution videos by employing convolutional neural networks (CNNs) with channel attention mechanisms (CAMs) and autoencoders (AEs). By enhancing blocks with more representative features, convolutional layers extract discriminating features from various networks. Additionally, we use random sampling of frames before main processing to improve accuracy while employing less data. The goal is to increase performance while overcoming challenges such as overfitting, computational complexity, and uncertainty by utilizing CNN-CAM and AE. Identifying patterns and features associated with selective high-level performance is the next step. To validate the method, low-resolution and low-size video frames were used in the UCF50, UCF101, and HMDB51 datasets. Additionally, the algorithm has relatively minimal computational complexity. Consequently, the proposed method performs satisfactorily compared to other similar methods. It has accuracy estimates of 77.29, 98.87, and 97.16%, respectively, for HMDB51, UCF50, and UCF101 datasets. These results indicate that the method can effectively classify human actions. Furthermore, the proposed method can be used as a processing model for low-resolution and low-size video frames.

1. Introduction

Human activity recognition includes a wide range of real-life applications, such as monitoring human activities, detecting abnormal or suspicious activity, retrieving video based on various actions, semantic video recognition, and observing patients in health centers [1, 2]. To date, several solutions have been proposed for monitoring actions using video images, such as a visual review of events in videos [3, 4]. While some have performed well, various body parts, such as hands and legs, can also be used to detect movement [5].

Still, images alone cannot depict the full action. Our ability to recognize complete actions in video data is based on analyzing human body movements in-frame and their interactions with the environment [6].

The system should function in a fraction of a second, which has unfortunately not received much attention in previous research. Although a compromise between accuracy and time is required, real-time processing is still regarded as one of the top benchmarks in information processing. Human activity recognition systems can process video frames based on frame rate per second and real-time

monitoring of nonstatic environments, according to statistics [7, 8]. It remains one of the most difficult aspects of video processing to track multiple goals in a chain of online videos. This is especially true when it comes to topics such as recognizing human activity. Databases contain movements in everyday life. These movements are considered normal, and some are considered anomalous [9–11]. Because of this, recognition under dense conditions is crucial in those multiple activities. In addition, accuracy is compromised when movements overlap, such as jumping and diving together. Therefore, we plan on developing an action recognition system based on a network of video sensors in different dynamic environments. This will apply to several multispectral control videos. In order to recognize data, feature extraction and classification algorithms are required, regardless of the type of data. Support vector machines (SVM) and neural networks (NNs) can be utilized as primary classifiers in handcrafted feature extraction-oriented systems like those described in [12]. Deep learning (DL), mainly convolutional neural networks (CNNs), based on the hierarchical system of the human visual cortex, has advanced considerably in image classification [13]. By using feature extraction and classification models, CNNs can learn categorical information from their features. Analyzing action representations and extracting features could significantly improve action recognition.

Human activity recognition is a challenging research field today. Video frames were analyzed to identify human activity. The demand for more precise and efficient frameworks for a variety of contexts grew, as did the demand for more information, images, and video frames. In this field, deep learning is a highly effective and powerful technique. Recently, several approaches have been presented to recognize human activity in video using CNN, also known as automatic methods. Nevertheless, such systems may not process multiple video frames accurately in real time. Consequently, the requirement for large volumes of real-time and offline data has led to creative ideas in the field of motion and activity recognition through video. Some general goals are as follows:

- (1) Our goal is to develop easy-to-use methods for our leading action recognition research. For various applications of human activity recognition, this is the most accurate estimate.
- (2) Our model has been trained and can be used in action recognition applications like hybrid deep learning. The network can thus extract information from several datasets and generalize it to other datasets, resulting in improved accuracy. Our model is, therefore, more efficient, faster, and more suitable for big data applications. The proposed model can be implemented as a ready deep learning architecture in action recognition applications due to its rapid convergence and updating.

Video processing should incorporate deep learning techniques, which uses several feature extraction models. CNN with autoencoder [11, 14, 15] (CNN-AE) characterizes

features well. CNN-AE extracts and classifies features based on improved attention mechanisms. As most methods for recognizing human actions rely on the quality of the frames, recognition errors may occur when the resolution or dimensions of the image change. Figure 1 illustrates how a decrease in quality can adversely affect recognition.

Despite the loss of some frame information, decreasing the size or resolution of video frames can have benefits when sending them to data centers. These benefits include preventing unnecessary operations like compression and decompression and online analysis of information received from the environment. Additionally, they reduce the complexity of computation. A suitable and fast structure, such as deep learning, can process low-size or low-resolution frames, reducing computation costs. DL-based structures can function in real time depending on how many layers they have. This reduces the decision-making component's computing complexity. As a result of the architecture proposed, it will be easier for an architecture based on generalizability, uncertainty, and evaluation criteria to be developed. A computational method for monitoring human activity is developed in this study using video frames of small size and a low number of frames. Smart city social systems can benefit from adopting and utilizing the proposed approach. To identify human actions in a video with increased accuracy, our research uses a hybrid structure combining a CNN structure and an AE network with a deep hybrid structure.

This study aims to improve human activity identification in video. Computational complexity is reduced by processing a small number of lower-resolution and lower-number frames in a short period of time. Our research is innovative in that it combines the improved CNN network with the channel attention module and the AE structure. This is for action recognition in high class numbers and in low-resolution videos. A structure like this has never been proposed in a similar study before.

This article considers the following contributions.

1.1. Generalizability and Robustness. The developed CNN with CAM and autoencoder (CNN-AE) model with attention mechanism (AM) is much more robust and helps the decision structure work more efficiently. The proposed method is considered robust since it has low dispersion and low accuracy against large frame quality changes. However, the diversity of datasets used and the ability of the method to make accurate decisions about unknown data demonstrate its generalizability. Due to its robustness, the proposed system recognizes human actions. On the other hand, the approach is capable of processing a random range of video frames of poor quality, indicating that it is sufficiently generalizable.

1.2. Monitoring Human Action. It is also possible to detect individual behavior. Monitoring unusual activities can serve many purposes. Recognition of human activity on video has a substantial impact on environmental deterrence and urban crime prevention, resulting in a more sustainable city.

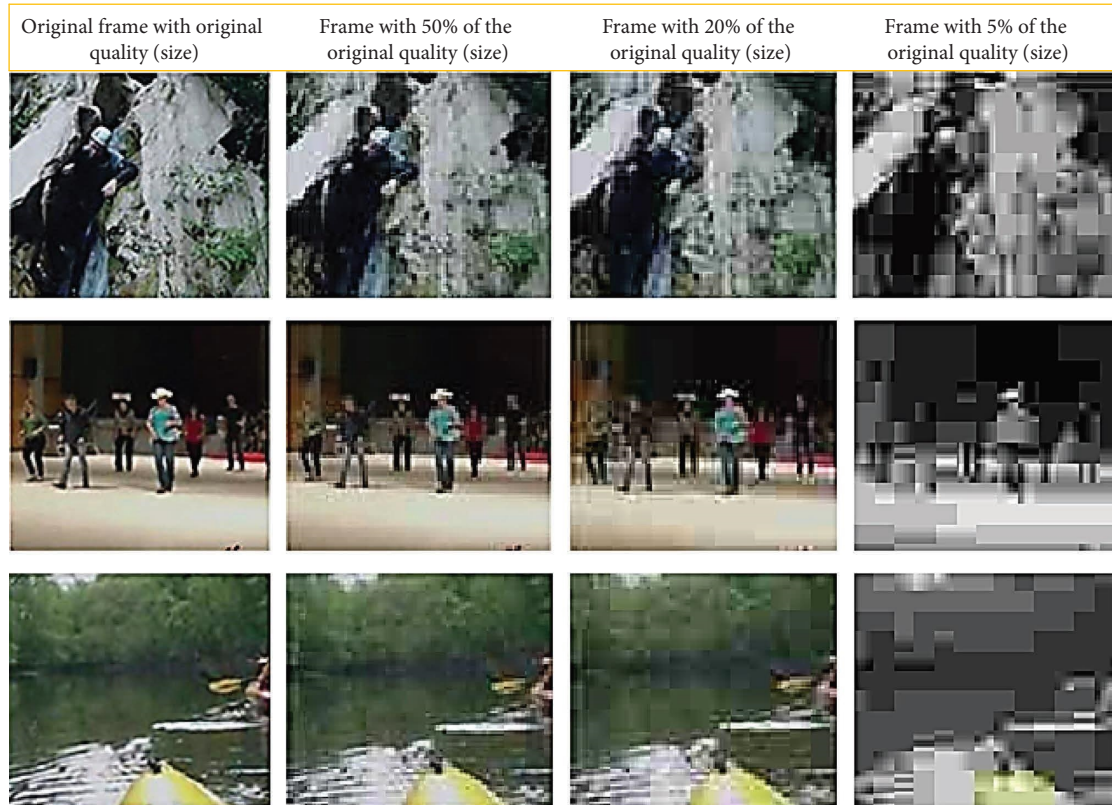


FIGURE 1: The recognition of actions in this figure is negatively affected by a decrease in resolution and dimensions.

1.3. Real-Time Decision-Making. A CAM-CNN architecture with AE architecture retains reliable recognition even when no frames or low-size videos are present, unlike end-to-end (e2e) and traditional deep learning models. According to some experiments, the proposed method represents a real-time method. As a single-processor, it has a sufficient frame rate of frames per second (FPS), and each frame takes less than a second to process. It is estimated that the proposed action recognition method can work in real time or close to real time. In other words, a fast structure that can recognize human actions quickly is relied upon during the decision-making process to aid in the processing of videos with small sizes and a low number of frames.

Our research is described below. Section 2 provides a brief overview of related studies. A newly developed feature extraction and learning technique is presented in Section 3 that uses the optimized CNN structure and AE described in Section 3. Section 4 reports the experimental outcomes generated by the proposed video frame analysis method. Following the conclusion of this study is a summary of the major points discussed in Section 5.

2. Related Work

In recent years, computer vision has gained interest in video comprehension and action recognition. This is due to its potential applications, such as robots, autonomous driving, camera monitoring, and human behavior analysis. The earliest video sequence encoding techniques used

handcrafted features [16–25]. With its rich trajectory features, AR with increased trajectories [19, 26] achieved remarkable performance and has become one of the most popular hand-designed systems today. In this section, we discuss two significant topics: deep learning-based action recognition approaches and low-resolution activity recognition methods.

2.1. Deep Learning-Based Action Recognition Methods.

Deep learning architectures [27–30] contribute to action recognition improvements [31–36]. Deep learning frameworks have been developed using computing hardware, such as tensor processing units (TPUs), graphics processing units (GPUs), and large-scale datasets, for action recognition. Accordingly, models such as MFNet-ACTF [25], asymmetric 3D-CNN [33], dual-attention network and RGB difference [35], recurrent neural network (RNN) [36], and action sequence optimization and two-stream 3D [37] were proposed to recognize different actions in videos. End-to-end classifications of spatiotemporal representations are the primary focus of these techniques. To simulate the temporal relationship between corporate 2D CNN variables, long short-term memory (LSTM) [38] is proposed in the study [39]. Although 2D CNN is more effective than handcrafted methods, it cannot adapt to motion changes [40].

It is possible to divide the remaining techniques for video action recognition into two categories. To enhance their temporal modeling capabilities, the first group of models uses a conventional two-stream structure [18, 41]. Spatial 2D

CNNs learn semantic features from optical flow, while spatiotemporal 2D CNNs analyze motion content from video. The final predictions are determined by averaging the scores of the two streams trained simultaneously. Data combinations for spatiotemporal analysis were examined in studies [37, 42–45]. Using sparse frames from evenly divided video clips, spatiotemporal segment networks (TSNs) [46] capture long-range relationships. Dual-path methods require optical flow computations because they are time-consuming and storage-intensive. The proposed technique, however, can operate without an optical flow mode, which reduces network complexity.

3D-CNN-based systems and 2+1D CNN systems comprise the second group of action recognition algorithms. 3D convolutions were used for the first time to define spatial and temporal data simultaneously in C3D [47]. As part of I3D [48], 2D convolutional kernels are supposed to be stretched into 3D to capture spatiotemporal features. There are, however, many parameters involved in 3D-CNNs, which makes them not suitable for all applications. A variety of strategies have been adopted to manage the costly calculations of a 3D-CNN using the 2D+1D paradigm. By decomposing 3D convolution into a pseudo-3D convolutional block, pseudo-3D (P3D) [49] produces a pseudo-3D convolution. 3D convolution is factorized by $R(2+1)D$ [50] and S3D-G [51] to improve precision and reduce complexity. A relational module can be viewed as an alternative to pooling using a time relation network (TRN) [52]. A spatiotemporal shift module (TSM) [53] shifts a proportion of features along the temporal dimension, giving the network the performance of a 3D-CNN while maintaining the complexity of a 2D CNN. With nonlocal neural networks [54], it was possible to capture long-range temporal dependencies between video frames and be more efficient. A dual-path network with an interactive fusion of mid-level elements was used in SlowFast [55] to model spatiotemporal data at two distinct temporal rates. Using the knowledge distillation procedure, our method also approximates the spatiotemporal representation at the feature level. The spatiotemporal representation capacity and transferability of 2D CNN and 3D-CNN models were determined [56]. Action recognition effectiveness can be enhanced by maximizing selected frames via dynamic knowledge propagation [57]. Elastic semantic networks (Else-Net) [58] and memory attention networks (MAN) [59] have shown improvement in recognition precision in recent years.

Frame ordering has been discussed in several previous works [60–62]. While these previous efforts partially addressed some aspects of order prediction, their results only provided limited supervision, i.e., a binary label for in-order or out-of-order events [60, 61] or subclip-based order prediction [62]. Furthermore, there is no explicit technique to encourage the model to prioritize motion data over background data.

Transformer-based techniques [63] significantly improve accuracy while conserving processing power. Using ViViT, a pure-transformer method for factorizing space-time dimension inputs, we handled spatiotemporal tokens from a long series of frames effectively. By separating spatial

and temporal focus within each block, TimeSformer [64] minimizes training time while maintaining test effectiveness. Spatial-temporal transformer (ST-TR) networks were constructed for skeleton-based action identification [65, 66]. In comparison with previous state-of-the-arts, Trear [67] has shown a significant improvement in egocentric RGB-D action recognition. Multiscale pyramid networks, MViT, were presented in [68] to extract information from low-level to high-level attention. Comparatively to other successful applications, transformers have not fully realized their potential in action recognition.

The human action recognition method has been employed for abnormal events and abnormal behaviors in some studies [69–72]. Additionally, it enhances safety and security by monitoring activities. Furthermore, it can be used to detect suspicious activity as part of a criminal investigation. Classical learning methods were used in some cases, while deep learning methods were utilized in others.

2.2. Low-Resolution-Based Action Recognition Methods.

Kawashima et al. [73] developed a deep learning-based method for identifying actions from extremely low-resolution thermal images. They distinguish between common and rare human actions (such as walking, sitting, and standing). Individual privacy protection is a strength of their work, which can be applied to Internet of Things (IoT) platforms. Low-resolution thermal images are difficult to compute feature points and build a precise contour of the human body, even if privacy concerns are overlooked. Thermal images, their frame differences, and the center of gravity of people's areas are used as inputs to their deep learning method for learning the spatiotemporal representation.

The application of deep neural networks to video action recognition follows their widespread adoption for image classification [47, 48, 74]. According to C3D [47], one of the most well-known deep networks, 3D convolution is more suited to extracting spatiotemporal features from video. Analysis of deep ResNet [27] structure options for action recognition [74] has demonstrated desirable performance on common benchmarks using I3D architecture [48]. The approaches to low-resolution (LR) single-frame applications include domain adaptation, feature learning, and super-resolution [48, 75].

Privacy protection has influenced earlier research on this topic [76–78]. The model in [77] identifies several transformations that produce LR videos based on the high-resolution (HR) training set. As a result of training on the LR dataset, action classifiers should gain a more precise decision boundary. The concept of inverse super-resolution (ISR) was introduced by Ryoo et al. [77] after they found distinct pixels in downsampled frames. Using this method, additional data can be extracted from low-resolution frames after learning how to alter images properly. To improve the acquisition of information inherent in low-resolution frames, Ryoo et al. [78] developed multi-Siamese loss. Ryoo's achievements have established the standard for recovering lost visual information from constrained pixels.

According to Chen et al. [79], LR and HR networks could share some filters in a semicoupled two-stream structure. It provides high-quality training frames. Xu et al. [80] found that leveraging HR videos effectively improved LR recognition performance significantly. A two-stream structure incorporating HR frames as inputs was demonstrated. A fully linked two-stream network that shares all convolutional filters with an LR network outperforms previous methods marginally. CNN-based action classifiers are trained simultaneously [79, 80] to ensure equal representation of HR and LR frames.

Action recognition [81] is examined in super-resolution. Optical flow-guided training was developed to improve existing image- and video-driven super-resolution architectures. They demonstrate their performance on genuine, minute actions by downsampling HMDB-51 and UCF-101 to 80×60 , but their performance on genuine, minute datasets differs greatly.

Novel models address the practical difficulties associated with extremely low-resolution activity [82–85]. Demir et al. [86] have also developed a natural LR benchmark called TinyVIRAT and an approach that employs a progressive generative method to enhance LR quality. By using these models in HR frames, visual information lost over time with a limited number of pixels can be retrieved [87].

Even though LR frames were used in most of these methods, it is unclear why more optimal architectures were not used. Conversely, similar methods have difficulty recognizing states like “falling,” “sitting,” and “lying down” because many action classes are not considered. Furthermore, some methods cannot be implemented in the real world as a model.

Despite the previous action recognition models, the paper presents an improved CNN that incorporates the structure with attention mechanisms and AE architecture. This will increase accuracy while using less information than previous models. In addition, we will test the method’s suitability for low-latency and real-time scenarios. Based on feature learning, we developed a dataset for short-term human action recognition using low-quality video. Similar action recognition models require scanning the entire length of a video sequence to classify large temporal sequences. Through this method, we can create a new and enhanced machine-learning tool for testing models that recognize human motions quickly and with minimal latency.

3. Methodology

Figure 2 illustrates how our model recognizes various actions in video frames using the introduced method. We describe this method in the following sections.

3.1. Preprocessing. In various environments captured on video surveillance, we use a deep learning network to recognize human actions and detect unusual activities or abnormal behavior. In addition to increasing accuracy, deep learning architectures are more capable of handling large datasets. Video input comes from a mix of existing and newly developed

sources. The process of preprocessing involves removing frames from previously captured videos. A subfolder named after each video is established and maintained along with the frames. JPG images are created from the video frames.

To conform to the enhanced integrated deep learning architecture, the data are compressed and saved in 224×224 dimensions. Prior to being stored in the folder, the testing video is also converted to frames and scaled to 224×224 . The preprocessing is performed using MATLAB functions. The bilinear method was also used for large, medium, and low-resolution or low-size images (i.e., 100, 50, and 10% of the original frame resolution). For downsizing images, a rapid reduction of dimensions or resolutions is preferred. Its bilinear frame downsizing accuracy and its speed are significant reasons for choosing it.

Random sampling is used to generate a few frames in an action video. By using frame sampling to reduce video volume, unnecessary data processing can be saved. Based on dataset characteristics, different videos have different numbers of shot segments. In order to reduce the number of images available for each segment, we randomly select one frame. Video captures almost all the actions with a small amount of information. As shown in Figure 3, we present a method for capturing dynamically sampled shots.

3.2. Proposed Hybrid Model. This paper describes a method for low-resolution action recognition and abnormal behavior from sample frames that consists of four sections: convolution, maximum integration, sampling, and fully connected. The following are parts of the proposed combined method to recognize human actions in video.

3.2.1. Multilayer Convolution. CNN architecture is depicted in Figure 4. Multilayer convolution has four types of operations: fully connected layer (gray color modules), up-sampling layer (light yellow modules), max-pooling layer (light green modules), and convolutional layer (light blue modules). The permeability of porous materials was predicted using a CNN (see Figure 4). There are two convolutional layers and one max-pooling layer in the CNN architecture. Max-pooling reduces the number of parameters in the network and expands its receptive field by halving the size of the feature map. As a result, the CNN structure is essentially the design of the network, while the autoencoder (AE) is the core of the network [14].

For AE and CNN, we provide frames of low-resolution $128 \times 128 \times 1$ size. The size of the detail matrix is reduced to $64 \times 64 \times 2$ after the first CNN layer. It is the number of kernels that determines the number of channels in the feature map when convolution is performed. Using the CNN architecture, a low-resolution $128 \times 128 \times 1$ -sized frame is converted to $4 \times 4 \times 32$, $8 \times 8 \times 16$, $16 \times 16 \times 8$, $32 \times 32 \times 4$, and $64 \times 64 \times 2$ -sized feature map. According to the most recent attribute map, each integer represents the highest level of a feature. To flatten and connect 3-D map layers, we used 1-dimensional feature lines with 512 features. AE creates a $4 \times 4 \times 64$ feature map, which is then transformed into a 1024-dimensional feature line. As shown in Figure 4,

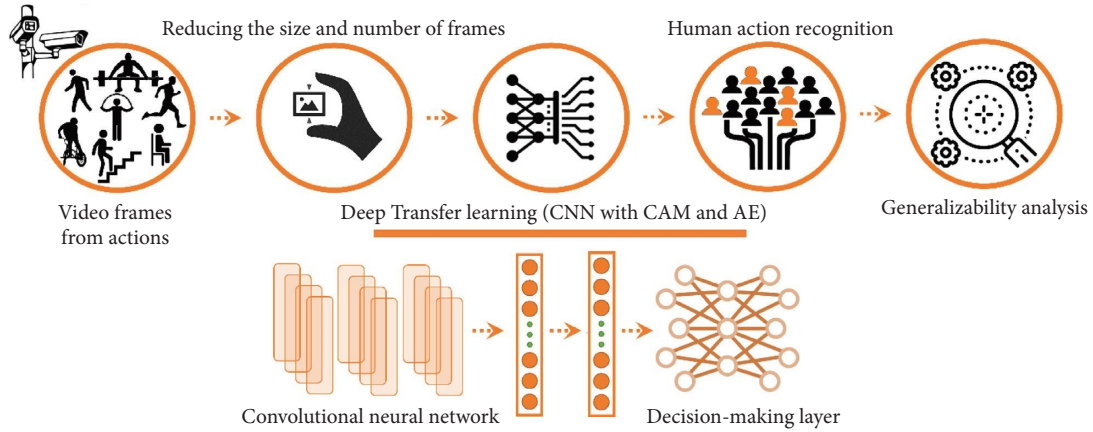


FIGURE 2: The stages of implementing the method for recognizing different actions in video frames.

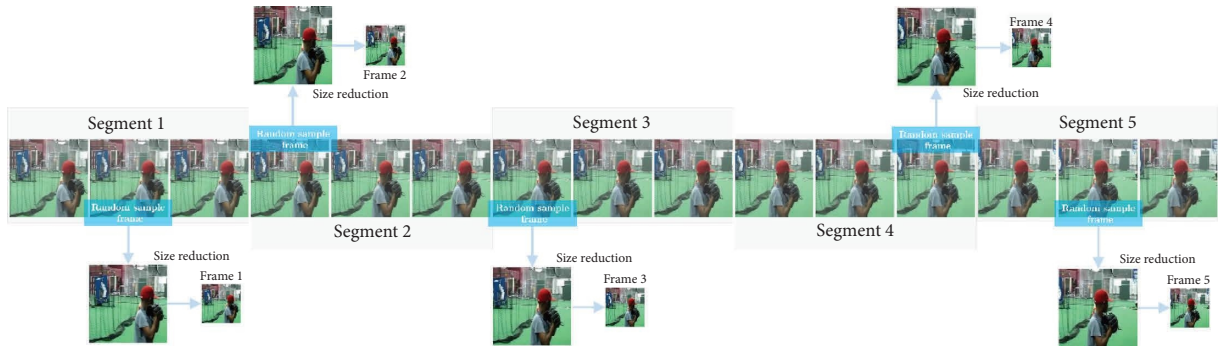


FIGURE 3: The random sampling for a baseball action. With the help of the downsizing strategy, a new and low-volume video is reconstructed by randomly choosing frames for each segment.

the AE-CNN will be discussed below. In addition, two feature maps are examined in an interconnected network. Input layers contain nodes that facilitate the transfer of low-resolution image output from one frame to the next.

Instances of a node may display regional characteristics, such as various parts of pixel picture information at different activity locations. Global characteristics can also be displayed in another instance. Training determines characteristics automatically. The fully connected network consists of nodes linked at the upper and lower layers. We use the nodes in the previous layer to calculate each node, which is expressed as follows:

$$x_{r,s} = b_{r,s} + \sum_i [x_{i,s-1} w_i]. \quad (1)$$

The current layer is indicated by s , the number of neurons in the layer by r , and the number of layers with full connectivity by w and b . A common machine learning strategy for evaluating, choosing, and utilizing high-level data to estimate valuations is a fully connected network. For instance, as depicted in Figure 4, it decreases in size from 400 to 150 due to classes. A frame can be used to deduce the actions to be taken in the upper half of the tree. If the input image has poor resolution, the reconstructed features will be inappropriate, common in feature engineering scenarios. Low-resolution frames lack comprehensive information,

resulting in confusion during training and accuracy drops. Low-level characteristics are needed to detect activities. CNN cannot forecast high-resolution properties based on low-resolution images. To support the trained network, low-resolution frames and high-resolution features can both be used. The hybrid CNN combines low-resolution images with features, while the AE module creates high-resolution images.

3.2.2. Autoencoder. To train AE procedures, we do not need to recognize every frame in the dataset. Relabeling, on the other hand, prevents low-detail frames from appearing and enables more accurate training. AE is significantly easier to collect training datasets due to labeled data independence. As a result, the dataset containing the greatest number of pairs of low- and high-resolution frames is selected as a starting point. The figure shows that the AE module contains an encoder (upper branch) and a decoder (lower branch). An encoder consists of three convolution layers and a max-pooling layer (distant branch). A decoder layer consists of one up-sampling layer and two convolution layers. Figure 4 illustrates in yellow how the aforementioned sampling approach has the opposite effect on the maximum collection operation. The small map is transformed into a large, high-resolution image using a sampling method that doubles its width and height. The encoder transforms low-

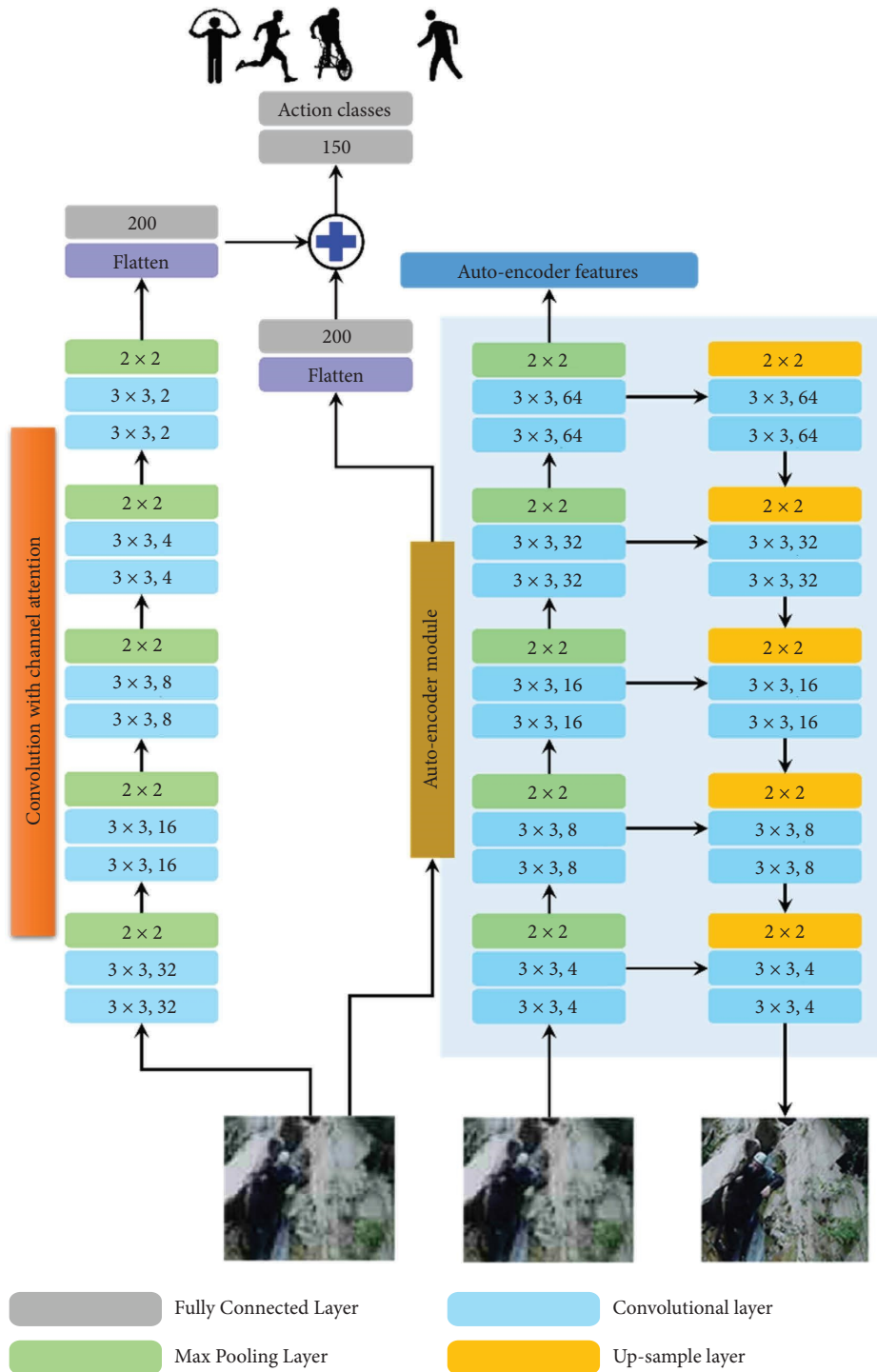


FIGURE 4: The framework of the proposed architecture is based on AE and CNN architecture. Network architecture has two branches. On the left side of the plot, a CNN is used with a channel attention block to recognize actions. Meanwhile, the right-side branch involves implementing an AE module to assess frame sequence characteristics.

resolution human activity data into a $128 \times 128 \times 1$ to $4 \times 4 \times 64$ -sized feature map using five-layer operation. In the decoder, human actions are simultaneously represented as high-resolution feature maps. By learning to infer high-resolution features from low-resolution images, the encoder passes the high-resolution intermediate variables to the original CNN to aid in prediction.

The training method becomes more challenging as layer variance increases with network depth. This obstacle was overcome by repeatedly training the AE module. $64 \times 64 \times 4$ indicates the initial map size, and each decoder and encoder have a layer allocated to it. The parameters of the layers are set, and the small network is trained. The encoder and decoder each receive another layer with convolution and

max-pooling layers, increasing the map size of the ultimate feature to $32 \times 32 \times 8$. This network's parameters have also been designed and trained. Through repetition, each encoder and decoder consist of five layers. The initial conditions are a low-resolution frame (L), a high-resolution or image (H), and a newly generated high-resolution or original-size image (newH). Four components of AE training are examined:

$$F = f(L) = \max_{\text{pooling}}(\text{ReLU}[\text{conv}(w_2, \text{ReLU}[\text{conv}(w_1, L)])]). \quad (2)$$

- (2) Unlike the previous step, the decoding procedure converts the F feature into a high-resolution newH

image. Input newF and output newH are related through the following equation [15]:

$$\text{newH} = f'(F) = \text{up}_{\text{sampling}}(\text{ReLU}[\text{conv}(w'_2, \text{ReLU}[\text{conv}(w'_1, L)])]). \quad (3)$$

The encoding and decoding convolution layers are identical with the exception of the last decoding layer. By improving the activation performance of the last complexity layer, the output result is transformed to the range 0-1.

- (3) The adaptive moment estimation technique reduces cross-entropy error for N data in AE (N_{AE}) by using a network that changes the network's settings [15].

$$\text{LOSS}_{\text{metric}} = N_{\text{AE}}^{-1} \times \left(\sum_i -[\log f_0(\text{newH}_i) \cdot H_i + \log f_0(1 - \text{newH}_i) \cdot (1 - \text{newH}_i)] \right). \quad (4)$$

- (4) During training, the number of encoding and decoding convolution layers increases. In both the encoder and the decoder, each layer is initialized one by one. Each encoder or decoder layer is added in three steps, up to five encoder layers.

$$\text{MSE} = \frac{\sum_i (y_i - y'_i)^2}{N}. \quad (5)$$

In this context, for N data, the variable y'_i represents the recognized action of the i -th low-resolution video image, while y_i represents the observed action of the corresponding high-resolution image as determined through the utilization of the lattice Boltzmann technique.

An encoder can achieve high-resolution recording of human actions by using the above training approach. This trains it to distinguish between low- and high-resolution frames from video frames. The decoder can produce high-quality images using this data. CNN's kernel was incorporated into an image processing module to extract features from low-resolution images. Both CNN and AE are provided as a fully connected layer for the ultimate prediction of actions from low-resolution images of distinct areas, with AE acting as a parallel branch line to the original CNN branch. Since the encoding features prevent deflection accumulation, we use them instead of high-resolution frames. For high-resolution frames, we need encoders and decoders before CNN, resulting in a 15-layer convolution layer instead of the 5-layer layer proposed in this study, which increases parameters, overfitting, and enhancement. Accuracy decreases when degradation occurs.

3.2.3. Channel Attention. Channel attention modules (CAMs) are CNN modules focused on channel-based attention. The channel attention map is generated by leveraging the interchannel relationship among features. The concept of channel attention arises from the understanding that each channel within a feature map detects specific features. Consequently, channel attention aims to determine the significance or relevance of the detected features in relation to the input frames. It is necessary to compress the spatial dimension of the input feature map to calculate channel attention effectively. A squeeze block and an excitation block were used in the feature channel domain. CNN extracts spatial features as a fitted decision system. By adjusting several feature maps in the channel domain, discriminating features can be selected.

In equation (4), the LOSS metric function is different from the loss function representing the entire combined network and its convergence. For this study, the LOSS function was used for the AE. However, in general, for the entire combined network and to guide the network to train all the parameters, the mean square error (MSE) was used as the loss function. The MSE can be expressed as follows:

Its performance can be maximized without adding new features by combining dense block and transmission layers with channel attention. Channel attention networks are small in size, and their assisting parameter is just 0.22 M, preventing overfitting. To minimize the size of the feature map, a transition layer with the 1×1 convolution layer and

a middle integration with stride 2 can be used. Combining the channel attention module with the transfer layer results in adaptive sampling. In Figure 5, the channel-based attention mechanism processes feature channels, such as “excitation” and “squeeze,” in two stages.

In the squeeze step, a one-dimensional vector of input characteristics is compressed into a length proportional to the number of input channels. In the original input feature, $W \times H \times C$, there are C channels in the spatial domain and U channels in the size domain.

The $1 \times 1 \times C$ vector is generated by compressing each spatial domain $W \times H$ into a single value by pooling global averages. The formal determination of the c_{th} component, z_c , of the squeeze output is given by the following equation [33, 59]:

$$z_c = F(u_c) = (H \times W)^{-1} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j). \quad (6)$$

Gate mechanisms consisting of two nonlinear, fully connected layers can capture channel dependence during the excitation phase.

As a result of the model’s low computational complexity, the two fully connected layers are just $C/16$ and C , respectively. s_c are used to represent the excitation output to decrease model complexity [33, 59]:

$$s_c = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)). \quad (7)$$

In the presence of W_1 and W_2 , which are the $C/16$ and C layer parameters, σ is the sigmoid function, and δ is the ReLU function. Furthermore, the z is the squeeze output. Finally, a weight is assigned to each feature channel. For each feature map, the weight vector s_c and the initial feature maps u_c are used as inputs. The channel-wise multiplication of feature maps produces the final product, the u'_c feature maps [33, 59].

$$u'_c(i, j) = u_c(i, j) \cdot s_c. \quad (8)$$

The channel attention module allocates adaptive weights to features by expanding and squeezing feature channels. The attention model for feature maps is the only parameter in this module that has a limited number of parameters.

4. Experimental Results

In this section, we analyze the results based on the implementation parts of the study methodology. We begin by examining the video frames.

4.1. Datasets. Datasets utilized in the analysis include HMDB51 [88], UCF50 [75], and UCF101 [76]. Dataset HMDB51 [70] is one of the most complex and difficult to analyze video image datasets related to human action recognition. Human facial interaction includes movement of body parts, physical contact with objects, and exercise. From YouTube, 6849 action samples were collected and categorized into 51 categories. Each category contains approximately 100 videos. Datasets are complicated when samples are collected from different participants performing the same task under different lighting and perspective settings.

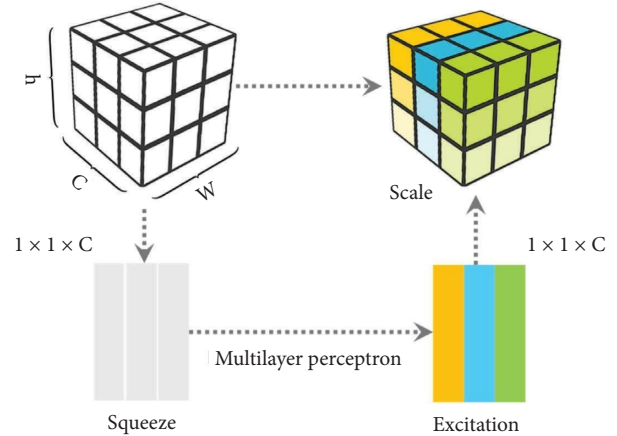


FIGURE 5: The channel-based attention mechanism by means of processing feature channels.

One of the most challenging datasets is the YouTube Action database. The action video images of people in this dataset are associated with low resolution, changing camera angles, changing scales, and bright and variable backgrounds. The dataset contains 11 sports classes with videos from 25 disciplines with four examples per action, as well as YouTube videos.

Considering the variety of camera movement, view and position of objects, object scale, perspective, cluttered background, and ambient light, the UCF50 [75] shows a wide range of human behaviors. The action groups are divided into several groups with some characteristics in common, such as a person who plays the piano four times from different perspectives.

It contains 13,320 YouTube videos from 101 action classes in AVI format from UCF101 [76]. Every action takes between 2 and 7 seconds, and 100 to 130 samples are evenly distributed across all categories. UCF101 analysis is difficult due to the large number of action classes involving human interaction with objects, musical instruments, and body parts. A few frames from the UCF50 dataset are depicted in Figure 6.

4.2. Implantation Details. The features of the computer system that allowed us to develop our approach are as follows: Intel (R), Core (TM), and Core i7 processors come with a single processor and 8 GB of RAM and a 64 bit operating system. MATLAB programming tools were used for the analysis of quantitative. The default learning rate for this model is 0.001. The improved model uses CNN and autoencoder between 200 and 1000 learning periods, and SGD applied CNN and autoencoder to further enhance the optimized structure. A single CPU processor was utilized to train the improved CNN model and autoencoder for about six to 10 hours for different learning structures.

All of our models are built based on transition learning models and fine-tuned convolutional networks. The training and validation process involved the calculation of errors, estimation of training parameters, convergence, and finally accuracy calculation. Error minimization during validation

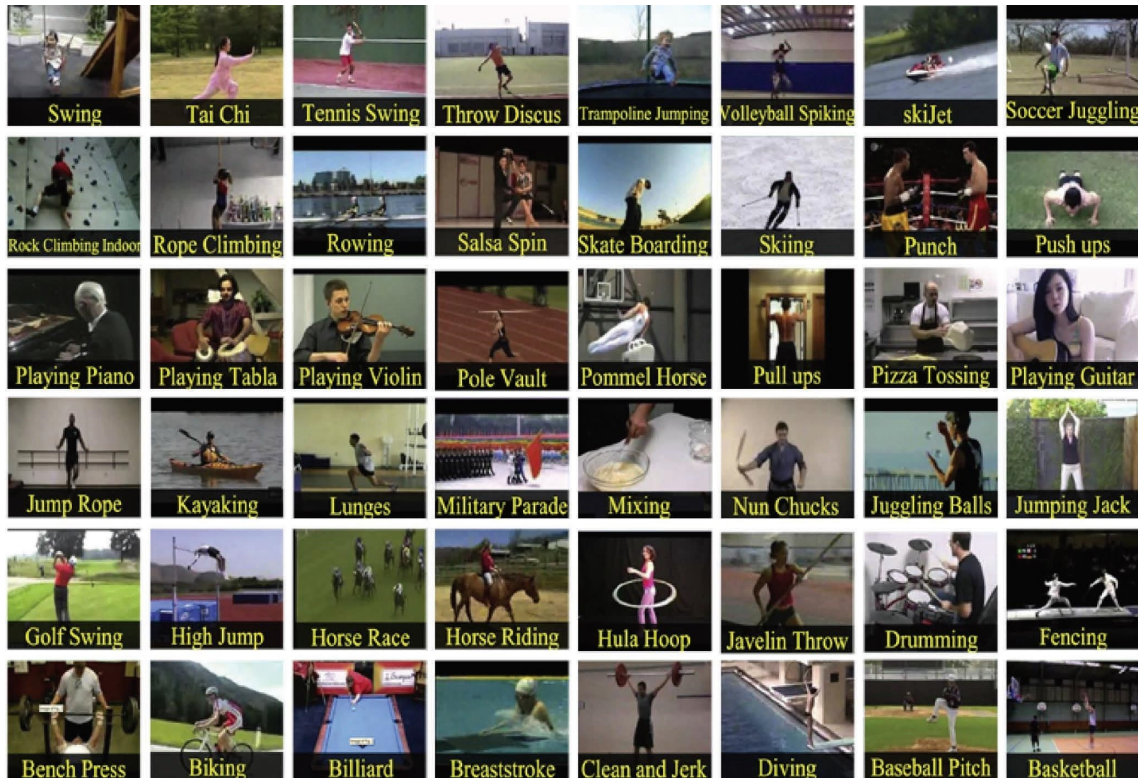


FIGURE 6: Several frames from the UCF50 dataset are depicted in this figure [75].

and training was used to identify the optimal convergence for each convolutional network structure. Insufficient progress in minimizing errors and checking accuracy automatically terminates the training process.

4.3. Evaluations. Based on the confusion matrix configuration, the multiclass status is estimated based on the accuracy criterion. In this study, three modes of all video frames were analyzed. In these modes, the frame was created at 70, 40, and 20% of the quality. The proposed model was used to identify human actions. An analysis of the confusion matrix determines how well a machine learning system performs in classification. The confusion matrix measures the difference between actual and expected values. Figures 7–9 show that the proposed method can recognize human activity at three different levels of video quality, i.e., 70, 40, and 20% of the original frames, with over 90% accuracy. It has even been observed that 100% accuracy has been achieved in some instances. There are separate sections for each assessment.

4.3.1. UCF50. As stated before, the films collected from this database are classified into 50 distinct categories. Each category's videos are broken into subcategories that share characteristics such as baseball, basketball shooting, bench press, and motorbike riding. Bicycling, shooting pool, diving, drumming, and numerous other activities are incorporated into sports. Some of them are quite similar to other human acts and movements. Figure 7 shows the algorithm results for three distinct video quality levels with

falling rosettes. While the frame size has not changed, the output accuracy varies slightly from the original resolution.

However, despite the drop in-frame resolution, the difference between the results is relatively small. The standard deviation is slight between them. Although the CAM-AE structure has a large number of classes, it has developed discriminative features and representation learning through changes in the set of frames.

As a result, the accuracy of more than 50 categories exceeded 96% and five of them exceeded 97% in the various action categories. Figure 10 shows the learning, training, and convergence process of the proposed method based on the model's accuracy and loss criterion. This is for the set of video frames obtained from UCF50 video data for all three types of frame quality. In comparison with other deep structures, the method identifies human actions with less computational complexity. The hybrid structure, however, will be more effective with more repetition. Moreover, there are also many layers of other CNN family structures with similar challenges, such as generalizability, uninterpretability, and computational complexity.

4.3.2. UCF101. The UCF101 dataset is complex and difficult to use since there are numerous action classes represented by humans who perform various activities with a variety of items, such as playing musical instruments, using sports equipment, or interacting with a procedure with different body parts. Figure 8 shows that when the size and resolution are reduced from 70% to 20% of the original frame, the classification error rate stays the same with low

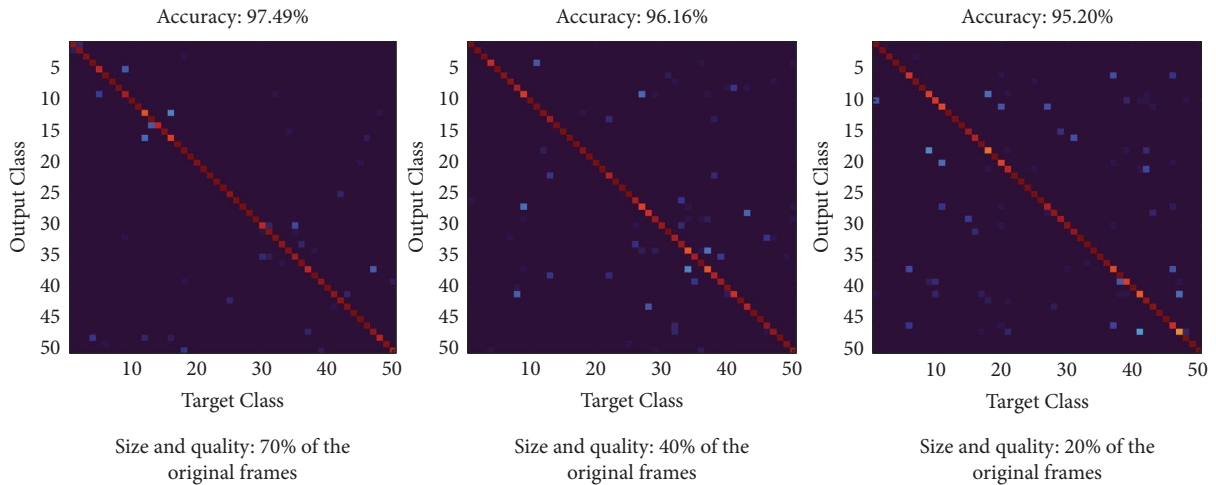


FIGURE 7: The confusion matrices of UCF50 action recognition datasets based on three different video quality levels.

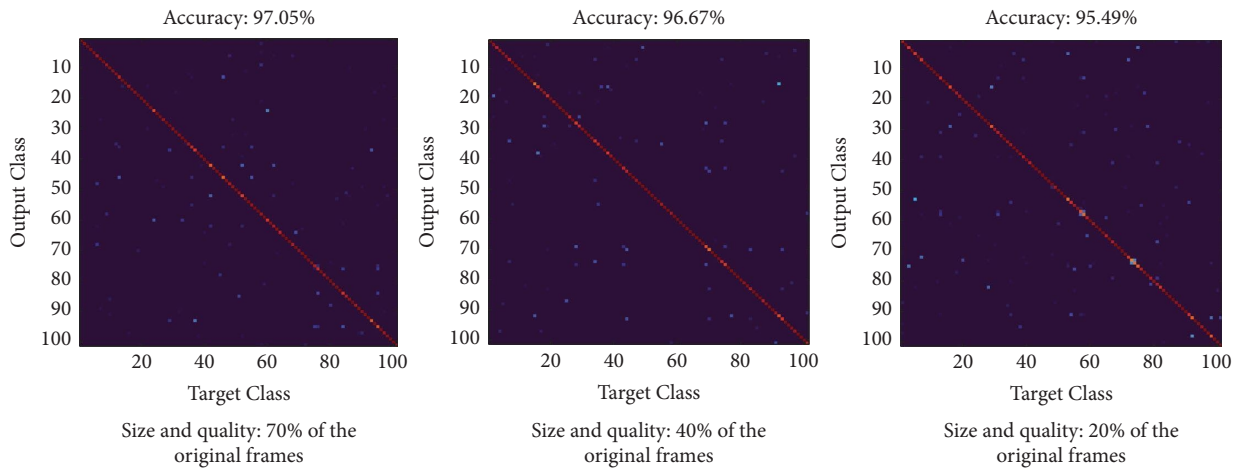


FIGURE 8: The confusion matrices of UCF101 action recognition datasets are based on three different video quality levels.

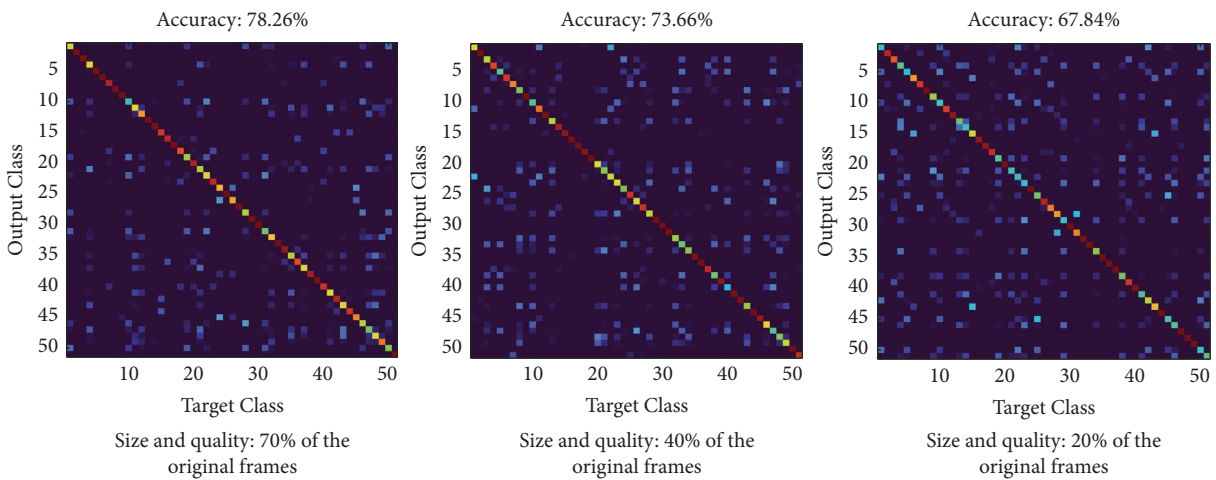


FIGURE 9: Based on three different video quality levels, the confusion matrices of the HMDB51 action recognition dataset are shown in this figure.

variances. Even when video frames are poor, processing has not been challenged and accuracy is higher than 96% in some cases.

For a set of UCF101 video frames with three different quality levels, Figure 11 illustrates the learning, training, and convergence processes of the proposed method. This is

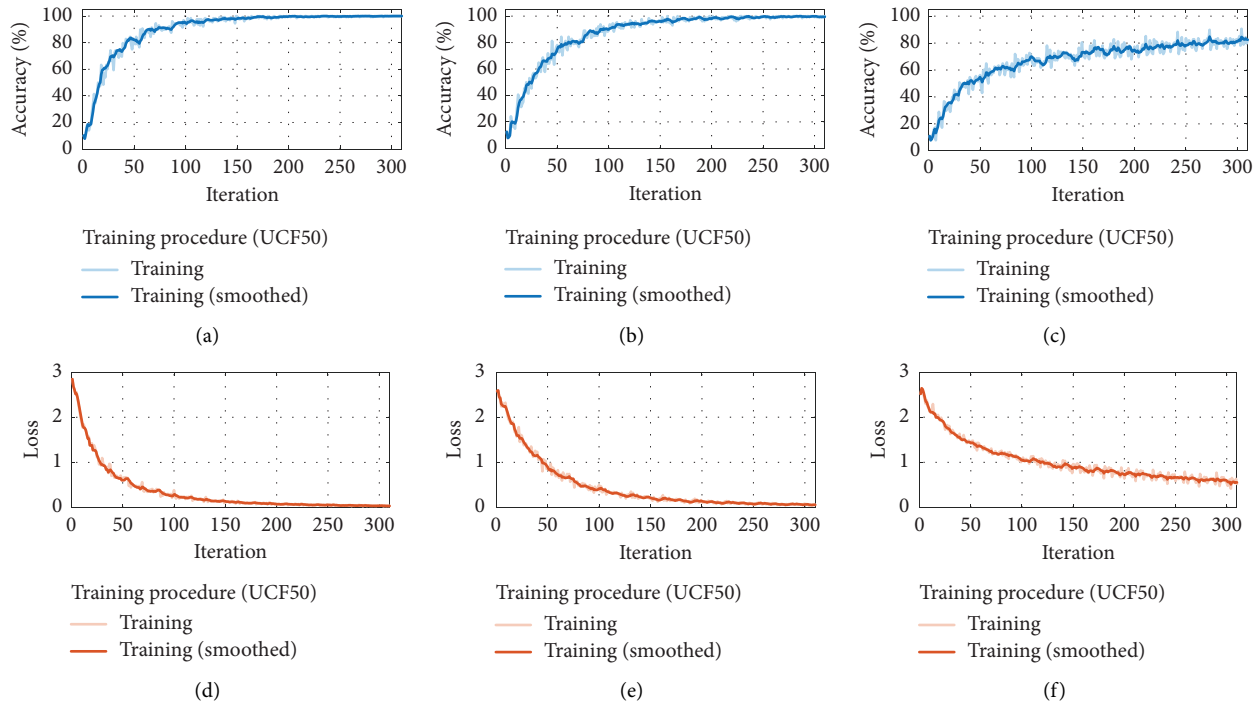


FIGURE 10: The training and convergence process of the proposed method is based on (a–c) accuracy and the (d–f) loss criterion of the model for the UCF50 video data for all three types of frame quality.

based on accuracy and loss functions. Moreover, it is evident that in addition to completing the claim of the previous section, the proposed method is more accurate and requires less computational complexity than other similar deep structures and algorithms for recognizing human actions.

4.3.3. HMDB51. The HMDB51 video frame set is one of the most complex sets of human activities ever studied. HMDB51 video frame set includes categories related to human exercise, body movements, and body contact with objects. In total, there are 6849 YouTube actions divided into 51 categories. There are approximately 100 videos in each category. Participants' varying brightness and perspectives have made the dataset more complex. State-of-the-art methods have 60% precision in this dataset. Interest in this form of data collection has grown dramatically in recent years, with some studies reporting a 70 percent interest rate. The suggested technique estimates a 78 percent increase in output despite video quality loss.

It is true that the proposed method for identifying human actions in the HMDB51 dataset is less accurate than that in the other two datasets; however, compared to other similar methods, the results are satisfactory. The obtained results are inaccurate due to the high complexity of the videos. There is little variation between reported outputs despite a significant quality drop. Figure 9 shows the results for three different video quality levels. Figure 12 depicts the training and convergence procedures of the proposed technique for a collection of HMDB51 video frames of three different quality levels.

5. Discussion

This research aims to reduce the number and size of video frames received from human actions while maintaining accuracy. Classification accuracy, however, will decrease as the video quality decreases. Through CAM and creating a deep hybrid structure with AE, the proposed method has overcome the challenge of low video quality in terms of frame number and size.

5.1. Recognition and Video Frame Quality. In Table 1, the performance of the proposed method is examined by reducing the dimensions of video frames as well as the number of frames. Labeling frames are determined by random sampling based on the original labels. To make the analysis less computationally complex, we randomly selected one of the three frames. When we analyzed what the final accuracy would be if a random frame were chosen from 2, 3, 4, ..., and 10 frames, we also considered other scenarios. Table 2 shows that the highest accuracy was obtained when one of the three frames was selected. Table 2 shows one frame at a time from 2, 3, 4, ..., and 10 frames.

In addition, frames with reduced dimensions were evaluated in terms of frames per second (FPS) to estimate the computational complexity of the frames. As the number of deleted frames increases, the correlation between the extracted features and the video sequence decreases. Methods may be used to dynamically find the most appropriate video frames. By using different strategies, preparing the video and finding the proper frames can, however, take a long time.

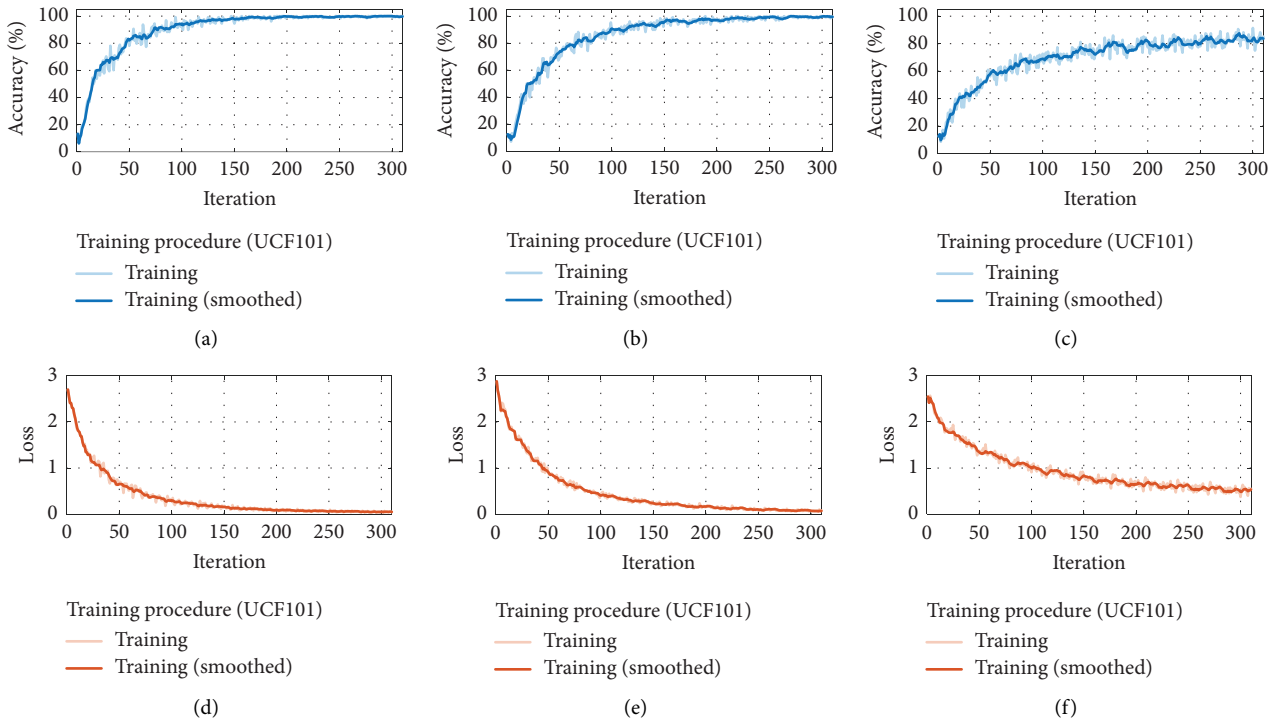


FIGURE 11: The proposed method is evaluated through accuracy evaluation (a–c) and loss analysis (d–f) in the UCF101 dataset for each of the three types of frame quality.

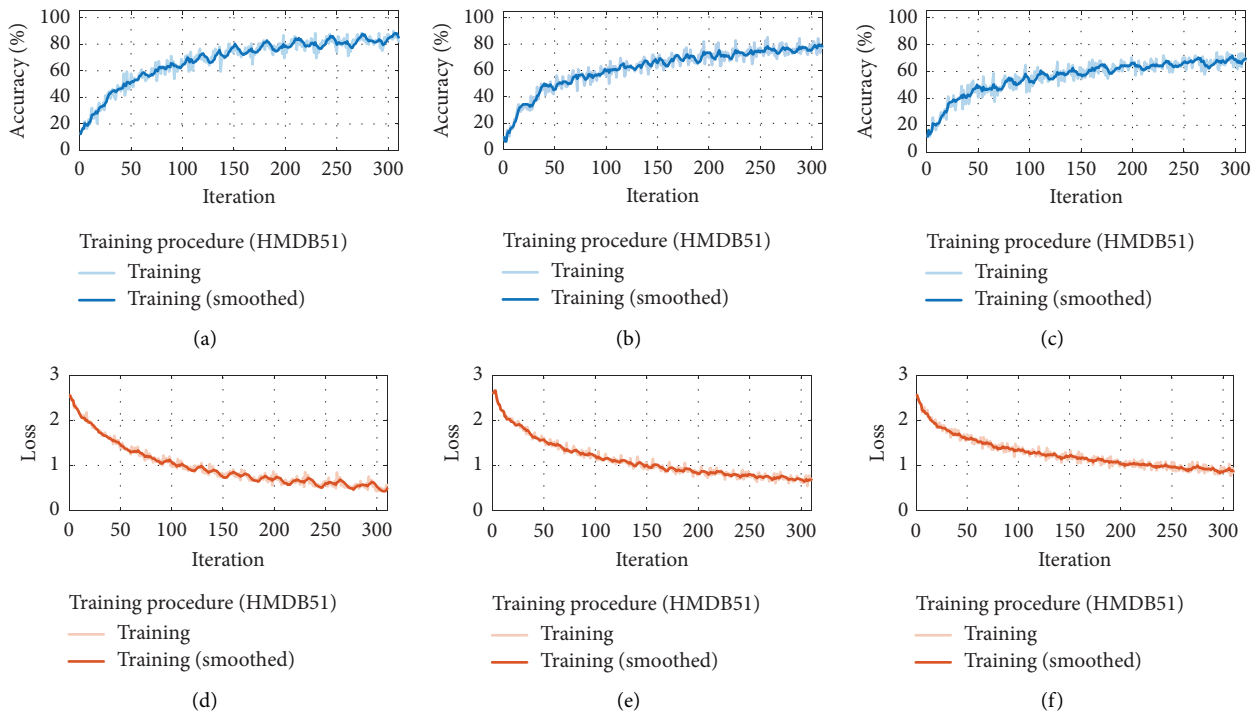


FIGURE 12: The proposed method is evaluated through accuracy evaluation (a–c) and loss analysis (d–f) in the HMDB51 dataset for each of the three types of frame quality.

5.2. Comparison. Comparative methods detect actions more accurately and with less computing expense than the suggested method. Using our method instead of handcrafted

methods, we extract features more accurately. Many new approaches to action recognition have appeared in recent years, including deep learning methods [20–37]. Despite

TABLE 1: By reducing the dimensions of the video frames as well as the number of frames, the accuracy of the proposed method is revealed in this table.



















| Qualities | | No. of frames | UCF50 (%) | UCF101 (%) | HMDB51 (%) |
|---|---|---------------|-----------|------------|------------|
| Dimensions | | | | | |
|  |  | | 98.44 | 98.13 | 85.76 |
|  |  | | 98.24 | 97.83 | 80.16 |
|  |  | | 97.87 | 97.16 | 77.29 |
|  |  | | 96.75 | 96.91 | 71.08 |
|  |  | | 96.39 | 96.58 | 70.33 |
|  |  | | 95.67 | 95.91 | 67.03 |
|  |  | | 95.03 | 95.51 | 66.88 |
|  |  | | 94.93 | 95.28 | 66.41 |
|  |  | | 94.80 | 94.91 | 65.21 |

TABLE 2: From several frame sequences, this table shows the accuracy, frame per second (FPS), and dimensions of choosing a frame.

| No. of selected frames | UCF50 | | | UCF101 | | | HMDB51 | | |
|------------------------|--------------|----------------|-------|--------------|----------------|-------|--------------|----------------|-------|
| | Accuracy (%) | Dimensions (%) | FPS | Accuracy (%) | Dimensions (%) | FPS | Accuracy (%) | Dimensions (%) | FPS |
| 1 from 2 | 98.23 | 100 | 11.53 | 97.93 | 100 | 12.08 | 85.28 | 100 | 12.79 |
| 1 from 3 | 98.11 | 80 | 14.87 | 97.58 | 80 | 13.91 | 83.71 | 80 | 15.03 |
| 1 from 4 | 97.41 | 70 | 16.34 | 96.87 | 70 | 15.63 | 80.69 | 70 | 17.61 |
| 1 from 5 | 97.12 | 60 | 18.90 | 96.03 | 60 | 18.56 | 77.42 | 60 | 19.30 |
| 1 from 6 | 96.79 | 50 | 21.44 | 95.55 | 50 | 20.93 | 74.37 | 50 | 22.15 |
| 1 from 7 | 95.30 | 40 | 24.74 | 94.10 | 40 | 24.17 | 74.37 | 40 | 23.87 |
| 1 from 8 | 93.84 | 30 | 26.38 | 92.89 | 30 | 26.12 | 70.61 | 30 | 25.64 |
| 1 from 9 | 92.50 | 25 | 29.11 | 91.17 | 25 | 28.85 | 68.78 | 25 | 28.81 |
| 1 from 10 | 91.73 | 20 | 30.49 | 90.43 | 20 | 30.24 | 66.14 | 20 | 29.13 |

their formidable structure, several methods based on deep learning and convolutional networks fail to overcome uncertainty obstacles by reducing the video's size, resolution, or frame number.

Besides extracting features, discriminating features that can be generalized under diverse acquisition conditions are essential. Feature extraction is sometimes achieved by creating a skeleton from the video; however, the information gained from the skeleton is sometimes discarded, making the method less robust and accurate [37, 45, 89, 90]. Several of the methods in [25, 32–36, 41, 45, 91–99] obstruct the operation of features by adding unnecessary parts. Thus, the addition will lower the accuracy of actions. Based on the attention mechanism, the autoencoder network, and the convolutional structure, the approach suggested in this paper has created a robust method that lowers video frame numbers and dimensions. The results are compared with those of similar approaches used in recent years as shown in Table 3. The method can also compete with deep learning-based methods that have emerged in recent years for action recognition [101–103].

For UCF50, UCF101, and HMDB51, the model learning training duration was 3, 5, and 4 hours, respectively. By using the benchmark dataset, the suggested classification approach is validated for its ability to achieve superior or comparable classification precision. We find that our suggested technique correctly detects human actions in videos in the majority of cases. Video information overlaps with human actions. Current approaches may incorrectly classify similar actions, such as drinking, eating, chewing, and talking.






5.3. Limitations. To date, considerable efforts have been dedicated to the recognition of human actions; however, only a limited subset of these efforts has adequately addressed the diverse range of limitations associated with this field. Video recording protocols for people's movements are one of the fundamental challenges encountered in this

domain. There are a variety of limitations involved, including time considerations, camera positioning, diverse weather conditions, video interference, and the inherent ambiguity surrounding movement classification. Human position and speed influence video images and recognition performance. As a result of excessive illumination and fluctuating weather conditions, human action recognition precision was occasionally compromised. A variety of camera angles make it difficult to accurately evaluate performance based on captured frames. Multiple instances of the proposed model's performance have been deemed satisfactory. However, it is still necessary to train it using videos. Complexity, duration, and poor quality of video frames are significant challenges in this task. It may be possible to conduct simultaneous activities over video. In contrast, humans engaging in multiple activities at the same time interfere with decision-making. It is necessary to consider distinct videos that can adequately train the model to address this concern. Human actions are intrinsically complex and challenging to comprehend. Additionally, most action recognition models on standard video datasets focus on videos captured under optimal conditions, ignoring videos captured under abnormal conditions. Moreover, implementation and constraint challenges may lead to pixel occlusion. Limitations such as camera movements and perspective distortion may influence individual actions. Recognition performance problems can be particularly aggravated when the camera moves. Variations in a system's operational classification affect its performance. There is a marginal difference between walking and running, for example. Understanding human behavior requires discernment between different categories. In scenarios involving changes in style, perspective, behavior patterns, and attire, recognizing human actions becomes increasingly challenging. Human-object communication and analogous activities remain active scholarly topics. In addition to monitoring and tracing multiple actions, recognizing irregularities, such as fraud detection and anomalous behavior, within a limited set of training data is challenging.

TABLE 3: Analyzing the proposed method against other comparable methods based on accuracy and computational complexity metrics.

| Method-year | Accuracy | | | Model | Feature extraction type | Computational complexity |
|------------------------|------------|-----------|------------|--|-------------------------|--------------------------|
| | HMDB51 (%) | UCF50 (%) | UCF101 (%) | | | |
| Liu et al. [20]-2015 | 49.95 | 75.60 | — | LLC and multiview pooling | Handcrafted | |
| Yang et al. [21]-2016 | 60.84 | 87.17 | — | Super-category exploration | Handcrafted | |
| Wang et al. [22]-2016 | 60.10 | 91.70 | 86.00 | HOG, HOF, and MBH with HD | Handcrafted | |
| Duta et al. [23]-2017 | 61.00 | 93.00 | 88.10 | HMG and iDT | Handcrafted | |
| Peng et al. [24]-2016 | 61.90 | 92.30 | 87.90 | Hybrid super vector and spatial-temporal pyramid | Handcrafted | |
| Xu et al. [25]-2021 | 76.3 | — | 96.3 | MFNet-ACTF | Handcrafted | |
| Huang et al. [26]-2017 | 63.30 | 93.30 | 88.10 | KNN-based trajectory pairing and STED | Handcrafted | |
| Duta et al. [31]-2016 | — | 95.40 | 91.40 | DA-VLAD two-stream | Deep features | |
| Sun et al. [32]-2017 | 66.20 | — | 93.60 | Lattice-LSTM | Deep features | |
| Shu et al. [33]-2018 | 46.01 | 93.75 | 76.07 | Open deep network | Deep features | |
| Hu et al. [34]-2019 | 90.80 | — | 88.20 | Space-time feature and spatiotemporal pyramid | Deep features | |
| Yang et al. [35]-2019 | 65.40 | — | 92.60 | Asymmetric 3D-CNN | Deep features | |
| Zhang et al. [36]-2020 | — | 60.40 | 91.00 | Dual-attention network and RGB difference | Deep features | |
| Xiong et al. [37]-2022 | 62.56 | 96.71 | 96.15 | Action sequences optimization and two-stream 3D | Deep features | |
| Deng et al. [41]-2021 | 71.3 | — | 95.3 | Diverse features fusion network (DFFN) | Deep features | |
| Wang et al. [45]-2021 | 74.1 | — | 94.5 | Spatial-temporal preference | Deep features | |
| Xu et al. [80]-2018 | 44.96 | — | — | w/o pretrained C3D | Deep features | |

TABLE 3: Continued.

| Method-year | Accuracy | | | Model | Feature extraction type | Computational complexity |
|------------------------|--------------|--------------|--------------|--|-------------------------|--|
| | HMDB51 (%) | UCF50 (%) | UCF101 (%) | | | |
| Zhang et al. [81]-2019 | 68.3 | — | 92.13 | Spatial-temporal ResNet | Deep features |  |
| Demir et al. [86]-2021 | — | — | 82.87 | TinyVIRAT | Deep features |  |
| Hou et al. [87]-2021 | 54.4 | — | — | Super-resolution-oriented generative adversarial network | Deep features |  |
| Zhao et al. [100]-2019 | 65.1 | — | 89.1 | Recurrent neural network (RNN) | Deep features |  |
| Proposed model | 77.29 | 97.87 | 97.16 | CNN with channel attention mechanisms and autoencoder | Deep features |  |

The best values are in bold.

6. Conclusion

Our method utilizes CNN-based channel attention mechanisms and autoencoders (AE) to recognize human actions in low volume and low number of frames dynamic video. Even low-quality videos transmitted over the Internet or from social media can be handled by our system. Additionally, CNN's model takes channel attention into account when choosing frame-level presentation. The designed AE can reliably identify multiple actions from poor-quality video frames. Before constructing a low-dimensional feature map, AE converts high-dimensional data into a low-dimensional feature map. Our experiments demonstrate that the proposed system is capable of processing a large number of frames per second (i.e., higher than 25 FPS) and can be employed in real time even when the resolution is poor. Using UCF50, UCF101, and HMDB51 benchmark datasets, this method identifies monitoring performance under nonstationary conditions. By using video frames with appropriate dependability ratings, the action recognition model can be fine-tuned to accommodate changes in nonstationary environments. With an improved version of our current system's architecture, our long-term strategy attempts to set and track specific goals. The video dataset does not include multiple actions performed by one individual. Actions that overlap, such as eating, drinking, and speaking, reduce video sample precision. As a multiview surveillance video architecture, we will develop a hybrid action recognition model. In addition, we will design a training architecture to overcome challenges such as noise, similar actions, actions under different weather conditions, and multiple actions at once.

Data Availability

Original publishers have allowed the authors to publish these images online without interfering with the data. All the data and codes are available through the corresponding authors.

Ethical Approval

Data used in this paper are publicly available and derived from a study by HMDB51 [88], UCF50 [75], and UCF101 [76], whose tests have been approved by the Ethics Board. In regard to their data, it is mentioned that the Organizational Ethics Board ignored the need for informed consent in reviewing these anonymous examples retrospectively.

Conflicts of Interest

All authors declare that there are no conflicts of interest.

Authors' Contributions

The experiments and data curation were designed and planned by E. Dastbaravardeh, M. Saberi Anari, and K. Rezaee. The formal analysis was carried out by S. Askarpour. The conceptualization of the proposed model was planned and implemented by K. Rezaee and

E. Dastbaravardeh. The visualization, writing, review, and editing were contributed by M. Saberi Anari, K. Rezaee, and S. Askarpour. Managing resources, implementing software, and supervising the project were the responsibilities of K. Rezaee. In addition to providing critical feedback, all authors assisted with editing the investigation and methodology. The manuscript was revised and edited by M. Saberi Anari and S. Askarpour.

Acknowledgments

Our sincere thanks go out to Meybod University and Islamic Azad University of Mashhad for their generous support. It would not have been possible for this study to be completed without the assistance of the Biomedical and Computer Engineering Department at Meybod University, Meybod.

References

- [1] W. Qi, N. Wang, H. Su, and A. Aliverti, "DCNN based human activity recognition framework with depth vision guiding," *Neurocomputing*, vol. 486, pp. 261–271, 2022.
- [2] G. Saleem, U. I. Bajwa, and R. H. Raza, "Toward human activity recognition: a survey," *Neural Computing and Applications*, vol. 35, no. 5, pp. 4145–4182, 2023.
- [3] X. Wang, S. Zheng, R. Yang et al., "Pedestrian attribute recognition: a survey," *Pattern Recognition*, vol. 121, Article ID 108220, 2022.
- [4] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: overview, challenges, and opportunities," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–40, 2021.
- [5] M. G. Morshed, T. Sultana, A. Alam, and Y. K. Lee, "Human action recognition: a taxonomy-based survey, updates, and opportunities," *Sensors*, vol. 23, no. 4, p. 2182, 2023.
- [6] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16979–16995, 2021.
- [7] A. Ullah, K. Muhammad, W. Ding, V. Palade, I. U. Haq, and S. W. Baik, "Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications," *Applied Soft Computing*, vol. 103, Article ID 107102, 2021.
- [8] R. Singh, A. Kumar Singh Kushwaha, R. Srivastava, and R. Srivastava, "Recent trends in human activity recognition—A comparative study," *Cognitive Systems Research*, vol. 77, pp. 30–44, 2023.
- [9] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Personal and Ubiquitous Computing*, pp. 1–17, 2021.
- [10] M. Hassaballah and A. I. Awad, *Deep Learning in Computer Vision: Principles and Applications*, CRC Press, Boca Raton, FL, USA, 2020.
- [11] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments," *Future Generation Computer Systems*, vol. 96, pp. 386–397, 2019.
- [12] O. Oreifej and Z. Liu, "Hon4d: histogram of oriented 4d normals for activity recognition from depth sequences," in

- Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 716–723, Portland, OR, USA, June 2013.
- [13] X. Yang and Y. Tian, “Super normal vector for activity recognition using depth sequences,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 804–811, Columbus, OH, USA, June 2014.
- [14] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, “Deep feature learning for medical image analysis with convolutional autoencoder neural network,” *IEEE Transactions on Big Data*, vol. 7, no. 4, pp. 750–758, 2021.
- [15] Y. N. Kunang, S. Nurmaini, D. Stiawan, and A. Zarkasi, “Automatic features extraction using autoencoder in intrusion detection system,” in *Proceedings of the 2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pp. 219–224, Pangkal, Indonesia, October 2018.
- [16] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *Proceedings of the 2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance*, pp. 65–72, Beijing, China, October 2005.
- [17] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,” in *Proceedings of the CVPR 2011*, pp. 3361–3368, Colorado Springs, CO, USA, June 2011.
- [18] S. Sadanand and J. J. Corso, “Action bank: a high-level representation of activity in video,” in *Proceedings of the 2012 IEEE Conference on computer vision and pattern recognition*, pp. 1234–1241, Providence, RI, USA, June 2012.
- [19] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in *Proceedings of the IEEE international conference on computer vision*, pp. 3551–3558, Sydney, Australia, December 2013.
- [20] J. Liu, Y. Huang, X. Peng, and L. Wang, “Multi-view descriptor mining via codeword net for action recognition,” in *Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP)*, pp. 793–797, Quebec City, Canada, September 2015.
- [21] Y. Yang, R. Liu, C. Deng, and X. Gao, “Multi-task human action recognition via exploring super-category,” *Signal Processing*, vol. 124, pp. 36–44, 2016.
- [22] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, “A robust and efficient video representation for action recognition,” *International Journal of Computer Vision*, vol. 119, no. 3, pp. 219–238, 2016.
- [23] I. C. Duta, J. R. R. Uijlings, B. Ionescu, K. Aizawa, A. G. Hauptmann, and N. Sebe, “Efficient human action recognition using histograms of motion gradients and VLAD with descriptor shape information,” *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 22445–22472, 2017.
- [24] X. Peng, L. Wang, X. Wang, and Y. Qiao, “Bag of visual words and fusion methods for action recognition: comprehensive study and good practice,” *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, 2016.
- [25] Y. Xu, J. Yang, K. Mao, J. Yin, and S. See, “Exploiting inter-frame regional correlation for efficient action recognition,” *Expert Systems with Applications*, vol. 178, Article ID 114829, 2021.
- [26] Q. Huang, S. Sun, and F. Wang, “A compact pairwise trajectory representation for action recognition,” in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1767–1771, New Orleans, LA, USA, March 2017.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.
- [28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 25, 2012.
- [29] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, <https://arxiv.org/abs/1409.1556>.
- [30] C. Szegedy, W. Liu, Y. Jia et al., “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, Boston, MA, USA, June 2015.
- [31] I. C. Duta, T. A. Nguyen, K. Aizawa, B. Ionescu, and N. Sebe, “Boosting VLAD with double assignment using deep features for action recognition in videos,” in *Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2210–2215, Cancun, Mexico, December 2016.
- [32] L. Sun, K. Jia, K. Chen, D. Y. Yeung, B. E. Shi, and S. Savarese, “Lattice long short-term memory for human action recognition,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2147–2156, October 2017.
- [33] Y. Shu, Y. Shi, Y. Wang, Y. Zou, Q. Yuan, and Y. Tian, “Odn: opening the deep network for open-set action recognition,” in *Proceedings of the 2018 IEEE international conference on multimedia and expo (ICME)*, pp. 1–6, Venice, Italy, July 2018.
- [34] H. Hu, Z. Liao, and X. Xiao, “Action recognition using multiple pooling strategies of CNN features,” *Neural Processing Letters*, vol. 50, no. 1, pp. 379–396, 2019.
- [35] H. Yang, C. Yuan, B. Li et al., “Asymmetric 3d convolutional neural networks for action recognition,” *Pattern Recognition*, vol. 85, pp. 1–12, 2019.
- [36] Z. Zhang, Z. Lv, C. Gan, and Q. Zhu, “Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions,” *Neurocomputing*, vol. 410, pp. 304–316, 2020.
- [37] X. Xiong, W. Min, Q. Han, Q. Wang, and C. Zha, “Action recognition using action sequences optimization and two-stream 3D dilated neural network,” *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, 2022.
- [38] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [39] J. Donahue, L. Anne Hendricks, S. Guadarrama et al., “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, Boston, MA, USA, June 2015.
- [40] T. D. Truong, Q. H. Bui, C. N. Duong et al., “Direcformer: a directed attention in transformer approach to robust action recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20030–20040, New Orleans, LA, USA, June 2022.
- [41] H. Deng, J. Kong, M. Jiang, and T. Liu, “Diverse features fusion network for video-based action recognition,” *Journal of Visual Communication and Image Representation*, vol. 77, Article ID 103121, 2021.

- [42] R. Christoph and F. A. Pinz, "Spatiotemporal residual networks for video action recognition," *Advances in Neural Information Processing Systems*, vol. 2, pp. 3468–3476, 2016.
- [43] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4768–4777, Honolulu, HI, USA, July 2017.
- [44] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933–1941, Las Vegas, NV, USA, June 2016.
- [45] J. Wang, Z. Shao, X. Huang, T. Lu, R. Zhang, and X. Lv, "Spatial-temporal pooling for action recognition in videos," *Neurocomputing*, vol. 451, pp. 265–278, 2021.
- [46] L. Wang, Y. Xiong, Z. Wang et al., "Temporal segment networks: towards good practices for deep action recognition," in *Proceedings of the European conference on computer vision*, pp. 20–36, Springer, Cham, Switzerland, October 2016.
- [47] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, Santiago, Chile, December 2015.
- [48] J. Carreira and A. Zisserman, "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, Honolulu, HI, USA, July 2017.
- [49] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, pp. 5533–5541, Beijing, China, December 2017.
- [50] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, Salt Lake City, UT, USA, June 2018.
- [51] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 305–321, Beijing, China, November 2018.
- [52] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 803–818, Piscataway, NJ, USA, June 2018.
- [53] J. Lin, C. Gan, and S. Han, "Tsm: temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7083–7093, Seoul, South Korea, October 2019.
- [54] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, Salt Lake City, UT, USA, June 2018.
- [55] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, Piscataway, NJ, USA, February 2019.
- [56] C. F. R. Chen, R. Panda, K. Ramakrishnan et al., "Deep analysis of cnn-based spatio-temporal representations for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6165–6175, Nashville, TN, USA, June 2021.
- [57] H. Kim, M. Jain, J. T. Lee, S. Yun, and F. Porikli, "Efficient action recognition via dynamic knowledge propagation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13719–13728, Montreal, Canada, October 2021.
- [58] T. Li, Q. Ke, H. Rahmani, R. E. Ho, H. Ding, and J. Liu, "Elsenet: elastic semantic network for continual action recognition from skeleton data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13434–13443, Montreal, Canada, October 2021.
- [59] C. Li, C. Xie, B. Zhang, J. Han, X. Zhen, and J. Chen, "Memory attention networks for skeleton-based action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4800–4814, 2022.
- [60] K. Hu, J. Shao, Y. Liu, B. Raj, M. Savvides, and Z. Shen, "Contrast and order representations for video self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7939–7949, Montreal, Canada, October 2021.
- [61] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *Proceedings of the European conference on computer vision*, pp. 527–544, Springer, Cham, Switzerland, October 2016.
- [62] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang, "Self-supervised spatiotemporal learning via video clip order prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10334–10343, Long Beach, CA, USA, June 2019.
- [63] K. Han, Y. Wang, H. Chen et al., "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023.
- [64] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proceedings of the 38th International Conference on Machine Learning*, vol. 2, no. 3, Piscataway, NJ, USA, July 2021.
- [65] C. Plizzari, M. Cannici, and M. Matteucci, "Spatial temporal transformer network for skeleton-based action recognition," in *Proceedings of the International Conference on Pattern Recognition*, pp. 694–701, Springer, Cham, Switzerland, January 2021.
- [66] Y. Zhang, B. Wu, W. Li, L. Duan, and C. Gan, "STST: spatial-temporal specialized transformer for skeleton-based action recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3229–3237, Beijing, China, October 2021.
- [67] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li, "Treat: transformer-based rgb-d egocentric action recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 1, pp. 246–252, 2022.
- [68] C. Feichtenhofer, "X3d: expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 203–213, Seattle, WA, USA, June 2020.
- [69] N. Aslam, P. K. Rai, and M. H. Kolekar, "A3N: attention-based adversarial autoencoder network for detecting anomalies in video sequence," *Journal of Visual Communication and Image Representation*, vol. 87, Article ID 103598, 2022.
- [70] B. Ramachandra, M. Jones, and R. Vatsavai, "Learning a distance function with a Siamese network to localize anomalies in videos," in *Proceedings of the IEEE/CVF Winter*

- Conference on Applications of Computer Vision*, pp. 2598–2607, Piscataway, NJ, USA, October 2020.
- [71] N. Aslam and M. H. Kolekar, “Unsupervised anomalous event detection in videos using spatio-temporal inter-fused autoencoder,” *Multimedia Tools and Applications*, vol. 81, no. 29, pp. 42457–42482, 2022.
- [72] N. Aslam and V. Sharma, “Foreground detection of moving object using Gaussian mixture model,” in *Proceedings of the 2017 International conference on communication and signal processing (ICCSP)*, pp. 1071–1074, Chennai, India, April 2017.
- [73] T. Kawashima, Y. Kawanishi, I. Ide et al., “Action recognition from extremely low-resolution thermal image sequence,” in *Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, Lecce, Italy, August 2017.
- [74] K. Hara, H. Kataoka, and Y. Satoh, “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, Salt Lake City, UT, USA, June 2018.
- [75] K. K. Reddy and M. Shah, “Recognizing 50 human action categories of web videos,” *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.
- [76] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: a dataset of 101 human actions classes from videos in the wild,” 2012, <https://arxiv.org/abs/1212.0402>.
- [77] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, “Privacy-preserving human activity recognition from extreme low resolution,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, CA, USA, February 2017.
- [78] M. Ryoo, K. Kim, and H. Yang, “Extreme low resolution activity recognition with multi-siamese embedding learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, California, CA, USA, April 2018.
- [79] J. Chen, J. Wu, J. Konrad, and P. Ishwar, “Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions,” in *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 139–147, Santa Rosa, CA, USA, March 2017.
- [80] M. Xu, A. Sharghi, X. Chen, and D. J. Crandall, “Fully-coupled two-stream spatiotemporal networks for extremely low resolution action recognition,” in *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1607–1615, Lake Tahoe, NV, USA, March 2018.
- [81] H. Zhang, D. Liu, and Z. Xiong, “Two-stream action recognition-oriented video super-resolution,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8799–8808, Seoul, South Korea, October 2019.
- [82] M. Monfort, C. Vondrick, A. Oliva et al., “Moments in time dataset: one million videos for event understanding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 502–508, 2020.
- [83] C. Gu, C. Sun, D. A. Ross et al., “Ava: a video dataset of spatio-temporally localized atomic visual actions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056, Piscataway, NJ, USA, April 2018.
- [84] S. Abu-El-Hajja, N. Kothari, J. Lee et al., “Youtube-8m: a large-scale video classification benchmark,” 2016, <https://arxiv.org/abs/1609.08675>.
- [85] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang, “Deep laplacian pyramid networks for fast and accurate super-resolution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 624–632, Honolulu, HI, USA, July 2017.
- [86] U. Demir, Y. S. Rawat, and M. Shah, “Tinyvirat: low-resolution video action recognition,” in *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 7387–7394, Milan, Italy, January 2021.
- [87] M. Hou, S. Liu, J. Zhou, Y. Zhang, and Z. Feng, “Extreme low-resolution activity recognition using a super-resolution-oriented generative adversarial network,” *Micromachines*, vol. 12, no. 6, p. 670, 2021.
- [88] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “HMDB: a large video database for human motion recognition,” in *Proceedings of the 2011 International conference on computer vision*, pp. 2556–2563, Barcelona, Spain, November 2011.
- [89] Y. Zhao, K. L. Man, J. Smith, K. Siddique, and S. U. Guan, “Improved two-stream model for human action recognition,” *EURASIP Journal on Image and Video Processing*, vol. 2020, no. 1, pp. 24–29, 2020.
- [90] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, “Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length,” *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 634–644, 2018.
- [91] X. Wang, L. Gao, J. Song, and H. Shen, “Beyond frame-level CNN: saliency-aware 3-D CNN with LSTM for video action recognition,” *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 510–514, 2017.
- [92] Y. Sun, X. Wu, W. Yu, and F. Yu, “Action recognition with motion map 3D network,” *Neurocomputing*, vol. 297, pp. 33–39, 2018.
- [93] G. Yao, T. Lei, J. Zhong, and P. Jiang, “Learning multi-temporal-scale deep information for action recognition,” *Applied Intelligence*, vol. 49, no. 6, pp. 2017–2029, 2019.
- [94] K. Liu, W. Liu, H. Ma, M. Tan, and C. Gan, “A real-time action representation with temporal encoding and deep compression,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 647–660, 2021.
- [95] M. Tong, K. Yan, L. Jin, X. Yue, and M. Li, “DM-CTSA: a discriminative multi-focused and complementary temporal/spatial attention framework for action recognition,” *Neural Computing and Applications*, vol. 33, no. 15, pp. 9375–9389, 2021.
- [96] Z. Li, K. Gavriluk, E. Gavves, M. Jain, and C. G. Snoek, “Videolstm convolves, attends and flows for action recognition,” *Computer Vision and Image Understanding*, vol. 166, pp. 41–50, 2018.
- [97] C. Li, B. Zhang, C. Chen et al., “Deep manifold structure transfer for action recognition,” *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4646–4658, 2019.
- [98] S. Rahimi, A. Aghagolzadeh, and M. Ezoji, “Human action recognition using double discriminative sparsity preserving projections and discriminant ridge-based classifier based on the GDWL-l1 graph,” *Expert Systems with Applications*, vol. 141, Article ID 112927, 2020.
- [99] C. Li, J. Zhang, and J. Yao, “Streamer action recognition in live video with spatial-temporal attention and deep dictionary learning,” *Neurocomputing*, vol. 453, pp. 383–392, 2021.
- [100] B. Zhao, X. Li, and X. Lu, “Property-constrained dual learning for video summarization,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 3989–4000, 2020.

- [101] Y. Li, G. Yang, Z. Su, S. Li, and Y. Wang, "Human activity recognition based on multienvironment sensor data," *Information Fusion*, vol. 91, pp. 47–63, 2023.
- [102] K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelwagen, "Delving deep into one-shot skeleton-based action recognition with diverse occlusions," *IEEE Transactions on Multimedia*, vol. 25, pp. 1489–1504, 2023.
- [103] S. Suh, V. F. Rey, and P. Lukowicz, "TASKED: transformer-based Adversarial learning for human activity recognition using wearable sensors via Self-Knowledge Distillation," *Knowledge-Based Systems*, vol. 260, Article ID 110143, 2023.