Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

Recognizing online video genres using ensemble deep convolutional learning for digital media service management

Yuwen Shao[1,2] and Na Guo[3*]

Abstract

It's evident that streaming services increasingly seek to automate the generation of film genres, a factor profoundly shaping a film's structure and target audience. Integrating a hybrid convolutional network into service management emerges as a valuable technique for discerning various video formats. This innovative approach not only categorizes video content but also facilitates personalized recommendations, content filtering, and targeted advertising. Given the tendency of films to blend elements from multiple genres, there is a growing demand for a real-time video classification system integrated with social media networks. Leveraging deep learning, we introduce a novel architecture for identifying and categorizing video film genres. Our approach utilizes an ensemble gated recurrent unit (ensGRU) neural network, effectively analyzing motion, spatial information, and temporal relationships. Additionally,w we present a sophisticated deep neural network incorporating the recommended GRU for video genre classification. The adoption of a dual-model strategy allows the network to capture robust video representations, leading to exceptional performance in multi-class movie classification. Evaluations conducted on well-known datasets, such as the LMTD dataset, consistently demonstrate the high performance of the proposed GRU model. This integrated model effectively extracts and learns features related to motion, spatial location, and temporal dynamics. Furthermore, the effectiveness of the proposed technique is validated using an engine block assembly dataset. Following the implementation of the enhanced architecture, the movie genre categorization system exhibits substantial improvements on the LMTD dataset, outperforming advanced models while requiring less computing power. With an impressive F1 score of 0.9102 and an accuracy rate of 94.4%, the recommended model consistently delivers outstanding results. Comparative evaluations underscore the accuracy and effectiveness of our proposed model in accurately identifying and classifying video genres, effectively extracting contextual information from video descriptors. Additionally, by integrating edge processing capabilities, our system achieves optimal real-time video processing and analysis, further enhancing its performance and relevance in dynamic media environments.

Keywords Video classification, Movie genres, Service management, Ensemble deep learning, Gated recurrent unit

Introduction

Visual media, such as images and videos, have become increasingly popular for sharing information due to their simplicity of creation and distribution [1–3]. The complexity of video perception has increased with the incorporation of temporal elements into spatial video, challenging image-based approaches. Creating a film from separate frames is straightforward, although the resulting quality is often subpar. Utilizing convolutional neural networks (CNNs) in

*Correspondence:
Na Guo
Guona2024@gmail.com
[1] China Tobacco Henan Industrial Co., Ltd. Technology Center, Zhengzhou 450000, China
[2] Zhengzhou University of Light Industry-College of Food and Bioengineering, Zhengzhou, China
[3] School of Arts and Education, Jinan Preschool Education College, Jinan 250307, China

computer vision has enhanced video categorization, which is essential for many computer vision tasks such as recommendation systems and video retrieval. Movie categorization is a challenging issue that requires more investigation, despite several studies conducted in this field [4].

Video-based movie classification has several applications such as genre-based video retrieval and filtering [5], automated labeling and annotation [6], and content recommendation [7, 8]. Early research focused on predicting film genres from input movies using domain-specific datasets as the primary aim [9–11]. Methods sometimes involve adapting models from existing video classification tasks related to the specific challenge, such as action identification or theme recognition. Several studies have employed shot sample methods to reduce computational costs [12–15]. The approaches included selecting sample images from whole films. These algorithms predict movie genres from posters or still frames, although they are constrained by dataset size and scalability. Researchers have explored methods to adjust models by utilizing large video datasets based on current information [7, 9, 11]. Challenges exist when implementing video classification systems for predicting movie genres. Prior research has typically overlooked linguistic cues in videos that might provide genre information [9, 10, 16–18]. Movie transcripts can be used to produce accurate predictions in some cases [19]. Sequences characterized by intense discourse and eerie music typically belong to the horror or thriller genre. In contrast, sequences utilizing upbeat language are usually found in the romantic or comedy genres. Furthermore, video classification frameworks often encrypt the entire movie to comprehend the content. Identifying and categorizing trailer genres illustrates video classification challenges. Trailers are sometimes classified by genre using phrases such as comedy or drama to indicate the themes of the films [20]. The technique of critics and IMDB users manually categorizing films is still common today, with the final decision made by the database. To optimize multimodal techniques, people are typically provided with both textual materials and films [21–23]. However, text materials require more effort, and their availability cannot be guaranteed at all times. From a sample of 33,000 movie clips taken from YouTube, Condensed Movies [24] found that 50% of them were without subtitles. The genres in this field of computer vision provide unique challenges because of their wide scope and overlap. This distinguishes them from other fields such as object tracking and activity detection [25]. Media outlets struggle to accurately represent genres. The only way to determine a genre is by seeing the entire film, not just parts of it. Films sometimes include aspects from many genres, making it challenging to classify them [26].

Personalized recommendations facilitate user engagement by tailoring material to individual preferences and needs. This promotes a deep sense of personalization and satisfaction, while also enhancing the likelihood that users will discover content they enjoy. Consequently, individuals are more inclined to engage in desired actions, such as reading, commenting, and sharing content. In addition, they are more inclined to increase their overall time spent on the website [27]. When a video is categorized as comedy, the viewer could be shown other relevant comical videos based on their interests [28]. This feature enhances customers' overall viewing experience by enabling them to explore and discover content according to their interests. Organizations can restrict access by using digital media service management, which ensures effective content filtering based on certain criteria. Managing digital media services significantly impacts targeted advertising efforts and enhances corporate efficiency and security. Organizations may significantly improve promotional activities by tailoring marketing campaigns to certain video genres using content filtering. Integrating digital media service management promotes tailored suggestions, resulting in a more immersive and customized user experience. We want to leverage the cross-correlation operator to develop a GRU neural network for movie genre classification that can effectively capture spatial and motion data and reveal their temporal links. A novel neural network resembling GRU is constructed utilizing convolution and cross-correlation methods. The distinctive GRU ensemble method acquires knowledge of temporal relationships and manages the extraction of spatial and motion data. Movie genres are classified using cross-correlation and convolution operators to extract motion and spatial data. We also explore spatial and temporal data upgrades to prevent overfitting. We conducted thorough testing on well-known video genre datasets to assess our proposed ensemble GRU's performance.

We collected an assembly dataset and utilized our framework GRU model to assess the effectiveness of our ensemble GRU architecture in engine manufacturing and upkeep. The dataset contains both commonalities and contrasts among classes, including a diverse spectrum of real-world challenges. The suggested approach demonstrates its ability to manage variations by replicating real production environments. Our ensemble-proposed GRU outperforms benchmark datasets, confirming our observed findings. Our work comprises three primary innovations. We introduce a novel ensemble GRU-like unit for video content analysis that simultaneously learns temporal connections and autonomously extracts spatial and motion information. We provide a deep learning structure for categorizing genres on the basis of our suggested GRU. The experimental results demonstrate that our developed GRU model is efficient and beneficial for classifying genres based on video frames. This study makes major contributions:

1) The incorporation of digital media service management is emphasized as a crucial element in improving efficiency and security in organizational environments. Furthermore, its expansion into targeted advertising through customized marketing campaigns in certain video genres is highlighted. This enhances the overall impact of promotional efforts by providing a more engaging and personalized user experience.
2) Our model presents an innovative method for creating an ensemble GRU neural network. The main goal is to use the cross-correlation operator to extract spatial and motion information, focusing on understanding their connection over time for classifying movie genres. This GRU-like neural network utilizes convolution and cross-correlation techniques to process spatial and motion information extraction while learning temporal relationships.
3) The paper describes comprehensive tests carried out on well-known video genre datasets to evaluate the effectiveness of the proposed ensemble GRU. The study involves using the suggested GRU structure in engine production and maintenance utilizing a compiled assembly dataset. The dataset includes several real-world issues such as similarities across classes and differences within classes, demonstrating the model's capacity to handle changes in actual production settings.
4) The achieved results of the proposed GRU ensemble on standard datasets demonstrate its superiority. Three main innovations in the work are highlighted: An innovative ensemble GRU-like unit has been introduced for video content analysis, along with an efficient deep learning architecture for genre classification using this GRU. The emphasis is on experimental results that highlight the functionality and effectiveness of the proposed GRU for genre classification from video frames.
5) Another significant contribution to research proposed in our work involves the active utilization of edge computing and cloud computing technologies in the process of identifying movie genres and enhancing digital media management services, as explored in this article. Recognizing that the analysis and categorization of videos necessitate substantial and scalable computational processing, leveraging scalable cloud computing resources offers a distinct advantage. Additionally, through the integration of edge computing, data can be processed in close proximity to their sources of production, thereby reducing latency and enhancing efficiency in video analysis. Consequently, the amalgamation of edge computing for localized computational processing and cloud computing for central computing in this research has resulted in performance and accuracy improvements in movie genre identification and the enhancement of digital media management services.

The remaining portion of the article follows this structure. "Related work" section offers a comprehensive exploration of the most recent and advanced techniques for classifying movie genres. A detailed elucidation of the proposed method is presented in "Proposed strategy" section. The dataset and experimental procedures are outlined in "Results" section. Serving as the conclusion to the article, "Conclusion" section delineates the subsequent steps and functions.

Related work

Automated movie genre categorization is a burgeoning field of interest, witnessing a surge in academic research. According to Rasheed et al. [29], video film genres can be determined by analyzing low-level statistics such as average shot time and color variation. Study [30] suggests the use of a neural network classifier for genre categorization with a single label, utilizing both visual and audio inputs.

Huang and Wang [31] employed a support vector machine (SVM) classifier for classifying visual and auditory input. Zhou et al. [32] proposed using picture descriptors like Gist, CENTRIST, and w-CENTRIST [33, 34], along with a K-nearest neighbor classifier for genre prediction. A Convolutional Neural Network (ConvNet) [35] can be utilized to generate image descriptions, enhancing genre prediction.

Multimodal approaches in video classification offer several advantages, capturing and analyzing audio, visual, and textual components to improve understanding and classification precision. Ogawa et al. [36] integrated multimodal learning, employing a bidirectional Long Short-Term Memory (LSTM) network for forward and backward dependencies in video data, with a classification strategy to distinguish favorite and non-favorite videos.

The incorporation of both spatial and temporal features simultaneously enhances video genre categorization. Alvarez et al. [37] emphasized fundamental movie elements, enhancing genre categorization outcomes. Ben et al. [38] utilized ResNet and SoundNet [39] to encode visual and auditory information, enhancing the assessment of temporal aspects with an LSTM network.

Yu et al. [40] proposed a bipartite model with attention, location, time, and sequence emphasis, utilizing a deep Convolutional Neural Network (CNN) for optimal movie frame results. Genre determination is achieved through a bi-LSTM attention model. Studies [41, 42] presents a probabilistic method, considering the importance of each background scene within different video categories, and recommends measuring shot length for varying genres.

Yadav and Vishwakarma [43] developed a system for classifying movie trailers by genre using deep neural networks. They trained a deep neural network with a large dataset of labeled movie trailers, achieving high accuracy

in genre categorization. Another study [40] employed convolutional neural networks to classify video game genres, predicting future game genres with labeled data and exploring methods like data augmentation and hyperparameter tuning for improved accuracy.

Studies [44, 45] categorized cinema genres using a multimodal method, combining visual features from trailers, textual features from synopses, and aural features from soundtracks. This amalgamation enhanced genre classification accuracy and reliability. Behrouzi et al. [46] integrated auditory and visual components through multimodal data, employing a recurrent neural network (RNN) for temporal correlations and achieving over 90% success rate in classifying different film genres.

There are two primary methods for categorizing movie genres: one relies on static imagery like posters and frames, while the other utilizes dynamic video content such as trailers and snippets. Recent research has shown a shift towards modifying existing frameworks for genre categorization in movies, departing from traditional video classification challenges. This adaptation includes integrating approaches to action recognition [16, 47–50] and video summarization [51, 52]. Many frameworks encounter challenges due to the high computational cost associated with video analysis. Utilizing methods that consider all frames as input for films longer than a few hours becomes impractical [49, 50]. Despite suggestions of sparse sampling methods to enhance efficiency, analyzing hour-long films still requires substantial computing resources. Automated movie genre categorization has gained significant traction in recent years, leading to a surge in academic research. Various methods and techniques have been proposed and investigated to address this task. Let's delve deeper into the methodologies discussed in the related work:

Rasheed et al. [29] propose analyzing low-level statistics such as average shot time and color variation to determine video film genres. This method offers simplicity in implementation and computational efficiency. However, it may suffer from lower accuracy due to its reliance on basic statistical features. Study [30] suggests employing neural network classifiers for genre categorization using both visual and audio inputs. Neural networks have the capability to learn complex patterns from data, potentially leading to higher accuracy in genre classification. However, they require large amounts of annotated data for training and can be computationally expensive. Huang and Wang [31] utilize SVM classifiers to classify visual and auditory input. SVMs are known for their effectiveness in handling high-dimensional data and can provide good classification performance. However, they may not perform well with large-scale datasets and require careful selection of kernel functions. Zhou et al. [32] propose using picture descriptors such as Gist, CENTRIST, and w-CENTRIST along with a K-nearest neighbor classifier for genre prediction. This approach leverages visual features extracted from images to classify genres. However, the performance heavily relies on the quality of the descriptors and the choice of the classification algorithm.

Multimodal methods, as highlighted by Ogawa et al. [36], integrate audio, visual, and textual components to improve genre classification precision. By considering multiple modalities, these approaches can capture richer information from the data, potentially leading to enhanced classification accuracy. However, integrating multiple modalities effectively can be challenging and may require sophisticated fusion techniques. Several studies, including those by Ben et al. [38] and Yu et al. [40], propose deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for genre classification. These models have demonstrated state-of-the-art performance in various tasks by automatically learning hierarchical representations from data. However, they often require large amounts of labeled data and substantial computational resources for training.

In response to the computational inefficiency observed in existing architectures and their inability to compete with models utilizing generalization techniques for enhanced decision-making processes, we propose a groundbreaking GRU-like neural network designed for the genre classification of a wide array of video films. This innovative network possesses the capability to internally and efficiently extract motion features through the strategic application of a cross-correlation operator. Additionally, to further enhance the real-time processing and analysis of video data, we integrate edge processing [53] capabilities into our proposed method. By leveraging edge processing at the network's periphery, we ensure swift and seamless processing, enabling more effective extraction of motion features and ultimately improving the accuracy of genre classification in real-time video analysis applications.

Proposed strategy

We present a comprehensive solution aimed at mitigating the prevalence of disturbing content in films. Harnessing the capabilities of deep learning, our innovative system has demonstrated significant efficacy in addressing the intricacies of video classification across diverse contexts. Our proposed methodology comprises three core elements: video preprocessing, deep feature extraction, and video representation and classification, as illustrated in Fig. 1. To ensure data integrity, rigorous preprocessing procedures are implemented to eliminate redundant or extraneous content from our dataset. Additionally, frames extracted from each video clip undergo standardization, ensuring uniform dimensions before input into convolutional blocks. These blocks, integrated into an ImageNet model, initialize the extraction of relevant features from each clip. Subsequently,

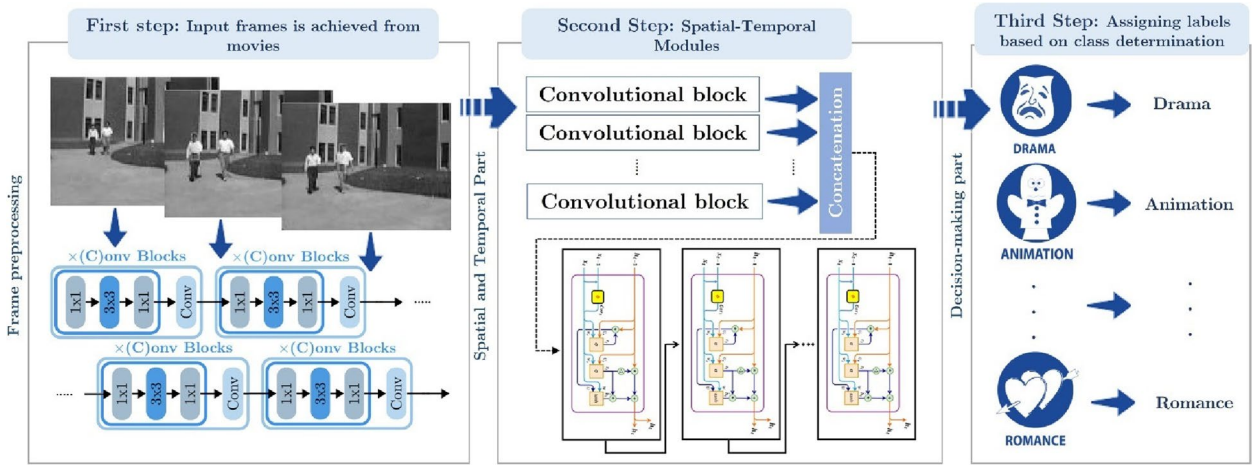

Fig. 1 In this figure, the proposed method for the genre classification of movies is shown

the extracted features undergo meticulous analysis within the ensGRU architecture to derive efficient and representative video representations, seamlessly integrated into the decision-making process. Each stage of our methodology is comprehensively detailed in the following sections.

Moreover, we present a proposed framework architecture designed to streamline the process of video genre identification. In the initial stage, videos undergo preprocessing to eliminate redundant frames and convert the remaining frames into distinct entities. Following this, convolutional blocks and ensGRU software are employed to extract feature vectors from the frames. The ensGRU architecture is then utilized to represent the video after transforming all feature vectors. The current procedure involves the use of a fully connected layer to compute the probability of a video clip being classified under a specific movie genre. Subsequently, an output layer incorporates a decision-making component.

Preprocessing step

Processing video frames is an essential step in various video analysis tasks, including classification, object detection, and action recognition. Preprocessing involves several operations aimed at enhancing the quality of video frames and extracting relevant information for subsequent analysis. The preprocessing pipeline typically commences with frame extraction, where consecutive frames are sampled from the video sequence at a fixed frame rate or keyframe intervals. This ensures a consistent representation of the video content and streamlines subsequent analysis. Once frames are extracted, they undergo various enhancement techniques to refine their quality and minimize noise. These techniques may involve operations such as denoising, contrast adjustment, and sharpening, aimed at improving image clarity and detail.

Following enhancement, we apply color normalization and standardization techniques to ensure uniform color representation across frames. This helps alleviate variations in lighting conditions and camera settings, thus enhancing the reliability of subsequent analysis algorithms. Additionally, spatial resizing and cropping may be employed to standardize the dimensions of video frames, rendering them suitable for input into deep learning models or other analysis algorithms requiring fixed-size inputs.

In addition to spatial preprocessing, temporal operations are also employed to capture motion information between consecutive frames. Techniques such as optical flow estimation are utilized to compute motion vectors between frames, providing valuable temporal context for tasks such as action recognition. The preprocessing of video frames plays a pivotal role in preparing raw video data for subsequent analysis tasks. By improving frame quality, reducing noise, and standardizing representation, preprocessing facilitates more precise and robust analysis of video content across various applications.

Convolutional blocks and concatenation

The "Convolutional Block-Concatenation" algorithm is an advanced method used to extract spatial features from videos for automatic genre classification. In this algorithm, convolutional blocks are used to extract important features from each video frame, and these features are then combined to form a comprehensive representation of the video. Since videos have variable lengths and the number of frames may differ, this algorithm helps us effectively extract features from all frames and arrive at an overall representation of the video. If the input frames are represented as X (video frame) and the output is denoted as Y (extracted features), then:

$$Y = Convolution\,(X) \tag{1}$$

In a Residual block (A type of pre-trained convolutional networks), if x is the input, the output is calculated as follows:

$$output = \mathrm{Re}\,LU(Conv(\mathrm{Re}\,LU(Conv(x)))) + x \tag{2}$$

where, Conv denotes the convolution operation, ReLU represents the rectified linear unit (ReLU) activation function, and "+" symbol is the direct addition operator between the input and the initial output. After extracting spatial features from each video frame using convolutional blocks, the extracted features from all frames are concatenated together to form a unified representation of the video. This concatenation is typically performed using distinct connections, where the features of each frame are merged in parallel. In the concatenation part, the input consists of y_1, y_2, ..., y_n, which represent the extracted features from all frames. Furthermore, the output is denoted as Z, which signifies the overall representation of the video. Lastly, the concatenation operation is defined as follows:

$$Z = Concatenate\,(Y_1, Y_2, ..., Y_n) \tag{3}$$

As outlined in Algorithm 1, the pseudocode outlines the process of convolution and concatenation, which involves extracting features from each frame of a video, and then concatenating those features to form the overall representation of the video. In addition to further analysis, the resulting video representation can be used to classify genres.

Ensemble GRUs

We propose that the Ensemble GRU model represents an advanced approach in the realm of deep learning, specifically tailored for video genre classification. This model is adept at processing sequential data types such as time series and videos. Leveraging the architecture of recurrent neural networks, it effectively captures dynamic relationships inherent in sequential data. Moreover, it employs the Ensemble technique, amalgamating multiple distinct GRU models, thereby enhancing accuracy and overall model performance. Typically, a GRU network comprises one or more layers of GRU units, with each layer sequentially processing input information and generating output. GRU units are equipped with gates that regulate information flow and facilitate long-term communication learning. The proposed Ensemble GRU model harnesses multiple independent GRU networks, amalgamating their outputs to formulate a final prediction. This approach commonly augments model accuracy and overall performance by capitalizing on the diverse learning capabilities of each GRU network. Collectively, these networks contribute to a more varied and comprehensive prediction. In the context of movie genre classification, each GRU network may discern distinct temporal patterns within movies, encompassing the progression of events, the frequency of emotional shifts, or the dynamics of character interactions. By consolidating the outputs of these networks, the Ensemble GRU model furnishes a refined and comprehensive prediction of movie genres.

Algorithm 1. Concatenation of extracted features from frames and pseudocode of convolutional blocks

```
# Input: video_frames (list of video frames)
# Output: video_representation (overall representation of the video)

# Create a list to store extracted features from each frame
features_list = []

# For each video frame
for frame in video_frames:
    # Extract spatial features using convolutional block
    features = convolutional_block(frame)
    # Append the extracted features to the feature list
    features_list.append(features)

# Concatenate the extracted features using distinct connections
video_representation = concatenate(features_list)

# Output: overall representation of the video
output = video_representation
```

Spatial–temporal module

While deep learning-based models are generally preferred over hand-crafted feature-based models, there is still significant work required to effectively learn various aspects of complex video data. Video data typically encompasses two fundamental features: motion information and spatial information, both of which evolve over time. Therefore, to comprehensively understand video data, it is necessary to address three primary sources of information: spatial information, motion information, and their temporal dependencies.

Existing models address these challenges by designing separate streams to extract spatial information using convolutional layers and optical flow using recurrent neural networks to capture temporal dependencies. However, unlike these models which often utilize mathematical methods or deep neural networks for optical flow extraction, our proposed approach employs a GRU neural network to simultaneously capture motion information through a spatio-temporal module. In our proposed model for video genre classification, as depicted in Fig. 1, the GRU neural network is designed to concurrently extract both spatial and motion features. This integrated approach allows for a more comprehensive understanding of the video content, leveraging both spatial and temporal information simultaneously.

GRU architecture

Initially, video frames are individually fed into two identical CNNs with shared weights. These CNNs are responsible for extracting spatial features from the video frames. Subsequently, the outputs of these CNNs are aggregated to produce a unified output. This aggregated output is then passed to a layer of spatio-temporal Gated Recurrent Unit that we have developed. In the final stage, the features obtained from the ensGRU layer are fed into a classifier.

Our model demonstrates the capability to capture both spatial–temporal properties and temporal relationships through the use of the ensGRU layer. Figure 2 provides an insight into the internal structure of the ensGRU layer. The ensGRU updating mechanism, as proposed, enables the extraction of motion characteristics and spatial information from videos while considering the temporal dependencies between these features. Figure 2 highlights the differences between our proposed ensGRU and the original GRU. Notably, in the ensGRU, inputs and weights are represented as 2D arrays, and convolution is employed instead of traditional multiplication.

Additionally, our proposed ensGRU can compute the correlation between two consecutive video frames by incorporating the video frame from the previous time step, t-1, as an additional input. The mathematical expression for the suggested ensGRU is as follows:

$$Corr_t = (x_t) \otimes (x_{t-1}) \tag{4}$$

Here, the Reset gate, Hidden state, New memory cell, and Update gate are defined as Eqs. (5) to (8):

$$r_t = \sigma_r(b_r + Corr_t * W_{rcf} + h_{t-1} * W_{rh} + x_t * W_{rx}) \tag{5}$$

$$h_t = (z_t \odot \widehat{h}_t) + ([1 - z_t] \odot h_{t-1}) \tag{6}$$

$$\widehat{h}_t = \tanh(b_{\widehat{h}} + Corr_t * W_{\widehat{hcf}} + (r_t \odot h_{t-1}) * W_{\widehat{hrh}} + x_t * W_{\widehat{xh}}) \tag{7}$$

$$z_t = \sigma_z(b_z + Corr_t * W_{zcf} + h_{t-1} * W_{zh} + x_t * W_{xz}) \tag{8}$$

The operator $\otimes$ represents the batch-wise correlation operator, while $\odot$ and $*$ represents the point-wise multiplication operators and convolutional function, respectively. The term "$Corr_t$" denotes the correlation matrix. To compute the correlation between x_{t-1}

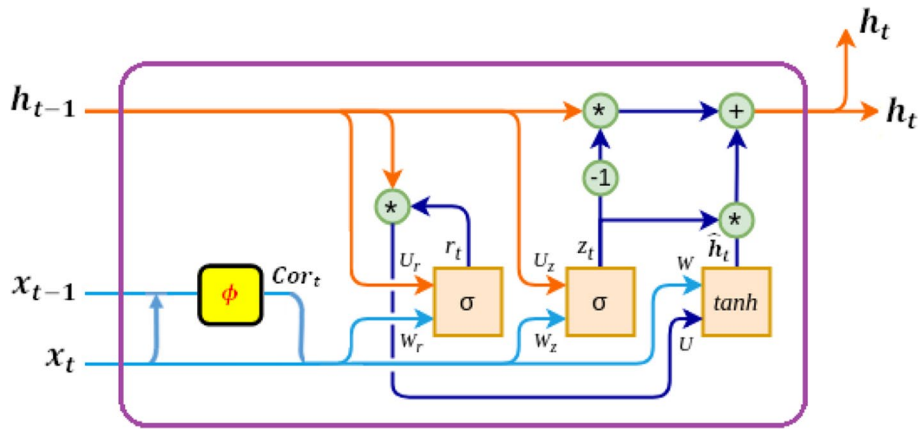

Fig. 2 This figure depicts the internal architecture of our proposed GRU

and x_t, the variables are partitioned into sub-matrices called patches. Subsequently, the normalized cross-correlation between these patches is computed using the following method:

$$L = N \otimes M \tag{9}$$

where,

$$L(x, y, z) = N(x, y) \circ M_z \tag{10}$$

which can be expressed based on (11):

$$L(x, y, z) = \sum_{j=-P}^{P} \sum_{i=-P}^{P} \frac{1}{\sigma_{M_z}} \frac{1}{\sigma_N} \left(-\bar{M}_z + M_z(i, j)\right)\left(-\bar{N} + N(x + i, y + j)\right) \tag{11}$$

where,

$$M_z = \left\{M(i, j) | (i, j) \in \{0, 1, 2, ..., w\} \times \left\{0, 1, 2, ..., h\right\}\right\} \tag{12}$$

and,

$$z = \left(\left[\frac{w}{l}\right] \cdot \left[\frac{j}{l}\right] + \left[\frac{i}{l}\right]\right) \quad s.t. \;\; l = 1 + 2P \tag{13}$$

The cross-correlation operation is represented by ∘. The collection of square patches M_z with sizes 1 + 2P is denoted as $M_{w \times h}$. To establish a stronger connection, we first normalize the input. Moreover, σ_{M_z} and σ_N are utilized for normalization, utilizing the standard deviation and mean of M_z, respectively. Since this normalization process does not carry any weight, it is not included in the training process.

Spatial and motion information can effectively be captured, and temporal interdependence can be modeled by employing both convolution and correlation methods. The cross-correlation procedure identifies portions of a picture closely resembling a smaller kernel image. Therefore, correlation is utilized to ensure that subsequent video frames exhibit similar patches. The heightened intensity in the central region of the right picture occurs when two photos display a high degree of similarity. The suggested ensGRU is employed to extract motion using the same approach.

The suggested correlation operator generates a meaningful 3-dimensional matrix, specifically referred to as ($Corr_t$ $Corr_t$). The model is trained by computing the values of $Corr_t$, x_t (the current input frame), and h_{t-1} (the output of the previous time step).

Edge computing

In our research, we have integrated edge computing technology to enhance the process of identifying online genres of videos, thereby improving digital media management services. In this method, video data sourced near production points, undergo processing via local or edge processing layers. To enact edge computing, we have devised a local processing system leveraging local hardware and edge servers [54, 55]. For instance, the implementation of edge computing technology enables the extraction of crucial features from videos at locations proximate to the source of video production, such as camera devices. This approach results in diminished latency and heightened accuracy in detecting key features pertinent to movie genres. Subsequently, by transmitting the extracted data to cloud computing environments, the process of classification and analysis is executed with greater precision. Cloud platforms boast potent and scalable computing resources, facilitating the execution of complex processing operations swiftly and with enhanced accuracy. This system comprises distinct processing layers operating concurrently:

a) Data Acquisition and Input: This layer entails the ingestion of video data from input cameras into the system.
b) Edge Processing: Within this layer, local image processing algorithms are applied to the input data. This step involves extracting key features from images and preprocessing them for subsequent use.
c) Transmission of Processed Data to Cloud Servers: Subsequently, the processed data from the edge servers are forwarded to cloud servers to facilitate a more accurate classification of video genres, leveraging the robust computing resources of the cloud.
d) Processing on Cloud Servers: In this layer, more intricate genre classification algorithms are applied to the video data. This entails the utilization of deep learning models to recognize video genres and categorize them accordingly.

By adopting this structure, the integration of edge computing enables a significant enhancement in processing speed and accuracy in identifying movie genres, while also efficiently distributing the computational workload between local resources and cloud servers. This optimized approach harnesses the computational prowess of both edge computing and cloud computing to yield superior and expedited results in the analysis and categorization of movie genres.

Results

We begin by providing precise details regarding the benchmark datasets utilized in our study. Following this, we will analyze the evaluation methodology. Subsequently, we delve into the specific details of the implementation. Finally, we compare our approach with the most exemplary ones in the field. The tables showcase the top performers in each metric by presenting their results in bold.

Dataset and setting

A total of 4021 videos were collected from the LMTD-9 multi-label trailer database [56] for the purpose of conducting this investigation. The diversity of trailer types found in LMTD-9 surpasses that of any other dataset available. The database consists of 603 sets, with 2815 sets allocated for testing and 603 sets for validation. The trailers analyzed in this study span across nine distinct genres, comprising 693 thrillers, 313 science fiction, 651 romances, 436 horror, 2032 dramas, 659 crime, 1562 comedy, 593 adventure, and 856 action alternatives.

To ensure non-overlapping frames, adjustments were made to the initial frame, facilitating the sampling of video clips at a rate of 22 frames per second. To accommodate video clips with frame rates outside the standard range of 23 to 24 frames per second, the method of padding and duplicating the final frame was employed. All experimental procedures regarding the training, validation, and testing of neural network models adhere strictly to a frame sampling rate of 22 frames per second (fps). The optimization of performance in feature extraction from video frames using an ImageNet pre-trained CNN model necessitates fine-tuning the hyperparameters of both the GRU and the CNN layers. Some examples of the videos from this dataset are depicted in Fig. 3. Although this dataset comprises multiple labels, certain researchers have focused on determining the dominant genre for each video and conducted subsequent analyses.

In the Convolutional Blocks, we employed the VGG-Net16 architecture, which comprises three essential components: the convolutional layer, utilized to extract spatial information. Each layer has been trained extensively on the ImageNet dataset for this purpose. A 3×3 matrix serves as the convolutional mask, selecting 3×3 patches to compute cross-correlation. Throughout this study, we investigate the Adam optimizer and softmax loss function for training purposes.

We extract a total of 30 frames from each video. Random stride and random starting points are utilized to sample each video clip. Various spatial data augmentation (DA) techniques, such as cropping, flipping, and rotating, are implemented to address overfitting concerns. Additionally, DA over time is employed to assemble a collection of films with similar themes but varying in pacing and genres. Moreover, $V \in R^{t\times w\times h\times ch}$ represents a color-channeled video comprising t frames, each with a size of $w \times h$. From the provided data, a sequence of video clips, denoted as $V_c \in R^{t\prime\times w\prime\times h\prime\times ch}$, $t\prime < t, w\prime < w,$, and $h\prime < h$, is extracted. In determining the selection of video clip $V_c = \left(f_{i_j}\right)_{j=1}^{t'}$ from the input video stream $V = \left(f_i\right)_{i=1}^{t}$, the following factors are taken into consideration:

$$V_c = \left\{f_{i_j} | i_s \in \left\{1, 2, 3, ..., t - t' \times s\right\}, i_j = i_{j-1} + s\right\} \tag{14}$$

Experimental results

To evaluate the accuracy and performance of the genre classification method, we first divide a set of movies into training, validation, and test sets. Then, we train the classification method on the training set and tune its parameters.

To assess the accuracy and performance of the method, we use various metrics such as accuracy (Acc), precision (Prec), recall (Recl), and F1 score (F1). After training the method on the training set, we evaluate it on the validation set and compare the results with the actual labels.

Additionally, after running the method on the test set, we compare its output with the actual labels and compute accuracy, precision, recall, and F1 score metrics for the model. These metrics help us evaluate the overall performance of the classification method in detecting and predicting movie genres and compare its accuracy and reliability. Table 1 illustrates a comparison of the test results obtained through our strategy.

The proposed methodology for categorizing cinema genres offers a viable alternative to conventional learning approaches such as VGG-16, VGG-19, and VGG-GRU (VGG-ensGRU). Utilizing a dataset comprising

Fig. 3 In this figure, examples of videos along with their corresponding genres from this dataset (LMTD-9) are depicted [56]

Table 1 The table provided presents our proposed methodology alongside experimental results from three similar experiments employing fivefold cross-validation studies using similar models. A diverse selection of videos was gathered from a wide range of films for analysis

Movie Genre	Strategy	Experiment 1				Experiment 2				Experiment 3			
		Acc	Prec	Recl	F1	Acc	Prec	Recl	F1	Acc	Prec	Recl	F1
Action	VGG-16	90.89	93.16	85.85	88.88	91.14	92.91	86.38	89.14	89.93	92.89	84.36	87.70
	VGG-19	92.18	94.85	87.40	90.48	92.82	94.91	89.03	91.53	92.16	93.79	89.90	90.37
	VGG-GRU	96.11	97.03	94.15	95.47	96.37	96.77	94.25	95.41	96.88	97.51	94.86	96.05
	VGG-ensGRU	**97.88**	**98.52**	**95.97**	**97.11**	**98.40**	**98.75**	**96.62**	**97.53**	**97.50**	**98.38**	**94.81**	**96.31**
Science fiction	VGG-16	89.35	91.99	83.32	86.79	91.02	92.84	85.38	88.29	91.54	93.67	86.51	89.40
	VGG-19	92.75	94.70	89.25	91.60	92.11	93.80	87.64	90.21	92.11	94.73	87.04	90.11
	VGG-GRU	94.23	96.12	90.96	93.15	94.80	96.10	94.41	93.45	93.39	95.59	88.71	91.49
	VGG-ensGRU	**95.26**	**96.38**	**92.11**	**93.93**	**95.96**	**97.33**	**93.30**	**95.09**	**95.32**	**96.42**	**92.44**	**94.22**
Adventure	VGG-16	91.98	94.54	87.03	90.08	92.17	93.76	87.52	90.14	92.05	94.06	87.29	90.08
	VGG-19	93.07	94.21	89.00	90.51	93.13	95.24	89.01	91.70	93.01	94.79	88.87	91.37
	VGG-GRU	94.61	96.45	90.63	93.04	94.23	96.34	90.23	92.74	93.59	94.71	89.31	91.62
	VGG-ensGRU	**94.87**	**96.23**	**92.13**	**93.98**	**95.32**	**96.80**	**91.73**	**93.80**	**94.68**	**96.00**	**91.28**	**93.29**
Comedies	VGG-16	92.82	94.11	88.80	90.99	92.75	94.84	88.24	91.03	93.27	94.88	89.20	91.60
	VGG-19	93.33	**95.22**	89.25	91.67	93.26	95.09	89.51	91.87	93.33	95.44	89.37	91.97
	VGG-GRU	94.03	**95.91**	90.78	93.02	96.35	95.50	89.45	91.94	94.17	95.63	89.97	92.40
	VGG-ensGRU	**95.19**	**96.04**	**91.48**	**93.44**	**96.15**	**97.13**	**93.56**	**95.15**	**95.64**	**96.97**	**92.84**	**94.68**
Crime	VGG-16	91.79	93.93	87.47	90.19	90.51	92.39	85.45	88.37	91.98	93.96	87.29	90.00
	VGG-19	92.37	94.93	87.63	90.59	92.69	94.07	88.74	91.03	92.17	94.10	87.94	90.57
	VGG-GRU	92.81	94.14	88.91	91.07	92.94	94.15	89.21	91.37	92.50	94.58	88.16	90.82
	VGG-ensGRU	**93.20**	**94.76**	**88.63**	**91.02**	**93.07**	**94.40**	**88.70**	**91.15**	**93.26**	**94.18**	**89.72**	**91.69**
Dramas	VGG-16	91.40	92.75	86.10	88.74	90.31	91.95	84.52	87.48	91.92	93.48	87.30	89.82
	VGG-19	93.20	94.97	89.80	92.04	93.20	94.63	89.33	91.59	93.07	95.66	88.65	91.57
	VGG-GRU	95.32	96.95	91.20	93.53	95.32	96.69	92.48	94.32	94.87	96.63	91.67	93.83
	VGG-ensGRU	**96.66**	**97.57**	**94.11**	**95.64**	**96.41**	**97.13**	**93.91**	**95.32**	**94.47**	**97.75**	**94.25**	**95.82**
Horrors	VGG-16	89.42	91.93	83.29	86.62	89.29	91.90	82.99	86.42	88.64	90.69	90.69	85.33
	VGG-19	90.94	92.80	85.50	88.46	90.64	93.48	85.51	88.65	90.64	93.06	85.58	88.66
	VGG-GRU	91.92	93.48	87.30	89.82	91.41	92.64	86.38	88.96	91.60	93.11	86.59	89.20
	VGG-ensGRU	**93.44**	**94.17**	**90.26**	**90.73**	**92.26**	**93.11**	**88.45**	**90.69**	**93.89**	**94.25**	**89.79**	**91.50**
Romances	VGG-16	89.93	92.52	84.20	87.54	89.69	92.36	83.38	86.71	90.44	93.26	85.57	88.75
	VGG-19	90.12	92.40	84.84	87.94	89.99	91.57	85.40	88.01	91.66	93.25	86.48	89.26
	VGG-GRU	91.73	93.97	86.98	89.90	92.50	94.20	87.60	90.28	92.82	93.39	88.83	91.00
	VGG-ensGRU	**93.97**	**95.39**	**90.30**	**92.47**	**93.97**	**95.26**	**90.35**	**92.45**	**93.53**	**95.23**	**88.70**	**91.20**
Thrillers	VGG-16	89.74	92.09	83.95	87.23	90.44	92.12	84.48	87.50	90.31	92.52	85.47	88.40
	VGG-19	91.28	92.67	86.86	89.35	91.20	93.49	86.18	89.19	91.03	93.15	85.85	88.84
	VGG-GRU	92.63	94.51	88.25	90.86	92.30	94.26	87.26	90.05	92.50	94.02	89.28	89.94
	VGG-ensGRU	**93.32**	**94.55**	**89.13**	**91.41**	**93.07**	**94.66**	**89.36**	**91.67**	**93.45**	**95.15**	**89.38**	**91.84**

preference samples from various video genres, ranging in magnitude from 1 to 9, our analysis reveals that this strategy is notably more effective and applicable across a broader range of scenarios than previously assumed. Achieving an impressive accuracy rate of 94.38% when tested on 9 distinct genres, our approach leverages video frames to cater to individual preferences, with each technique achieving the desired level of accuracy.

Results indicate that employing the fivefold cross-validation (CV) technique enhances experimental accuracy while substantially mitigating potential bias, yielding reliable outcomes. However, some elements within the video frame collection may exhibit suboptimal performance. To address this, three enhanced models—VGG-16, VGG-19, VGG-GRU, and the proposed model—were employed, as shown in Table 1. Our research underscores

these models' exceptional computational accuracy, reactivity, and ability to discern between different data.

Consequently, our design surpasses state-of-the-art techniques in both accuracy and computational cost. Based on trial data, VGG-GRU outperforms other frameworks with an F1 score of 0.9102 and an accuracy of 93.70%.

The categorization strategy yielded valuable data with a precision rate of 93%. The proposed approach demonstrated superior performance compared to similar models in a nine-class classification task, achieving an accuracy rate of 94.3% across all three rounds of fivefold cross-validation. Multi-class categorization was applied to each model scenario. Confusion matrices, as depicted in Fig. 4, serve as confusion matrixes for model comparison and representation of categorization outcomes. The data did not exhibit any statistically significant variation

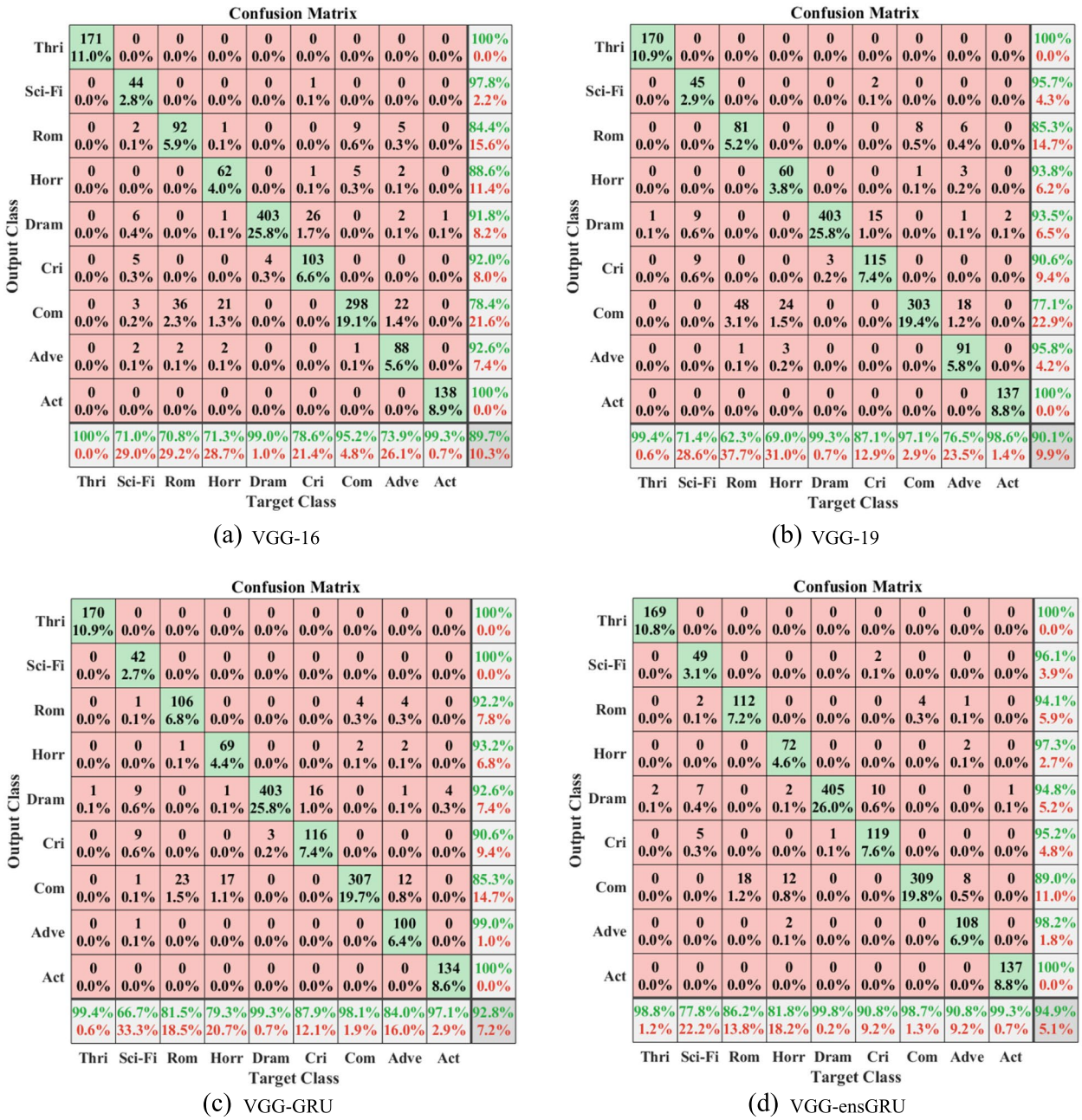

Fig. 4 The presented findings showcase confusion matrices that illustrate the classification outcomes of three comparable models in the context of movie genre categorization, in comparison with the proposed model

or deviation, with a categorization accuracy rate of 94.4% across various movie genres.

Given the subjective nature of categorization, films are often classified into a single genre rather than multiple genres. Evaluation of classification output and the approach's effectiveness across various movie genre classification scenarios is conducted using variance estimation. Figure 5 displays classification results from several tests, compared to similar models such as VGG, VGG-GRU, ensGRU, and VGG-ensVGG, using receiver operating characteristic (ROC) curves and area under the curve (AUC) for each class. Our proposed model consistently outperforms VGG-GRU, ensGRU, and CNN + ensVGG architectures in genre categorization, demonstrating exceptional classification accuracy and robustness in the face of ambiguity.

Discussion and comparison

The model we have designed, named ensGRU + CNN, has remarkable capabilities and is comparable to other methods such as VGG, ResNet, and DenseNet in video genre classification. Its accuracy for identifying 9 different genres in the LMDT dataset is higher than three other methods, at 94.4%. This model stands out from other methods due to its combination of GRU and CNN networks. One of its strengths is its high ability to detect temporal and

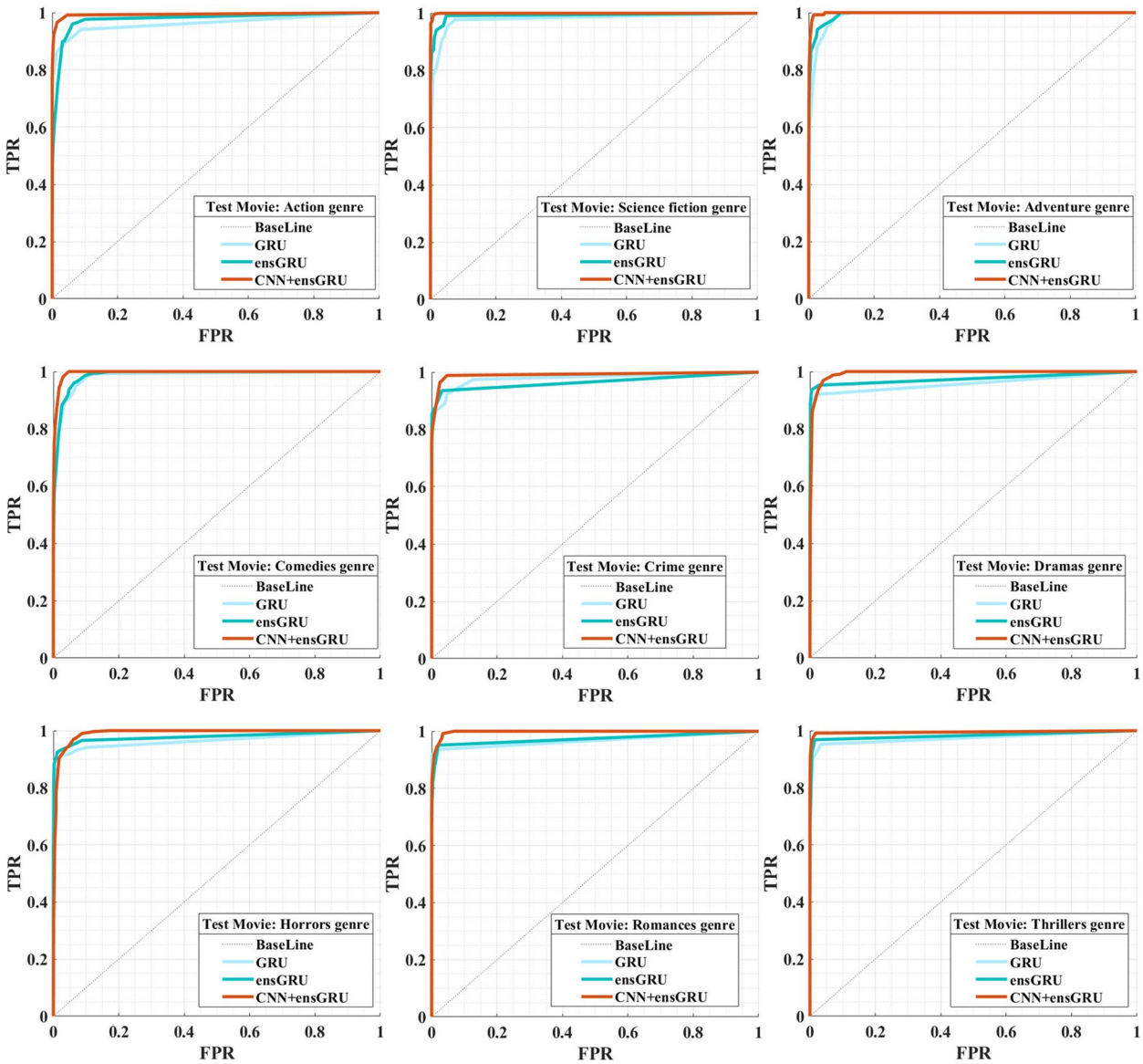

Fig. 5 An assessment has been conducted to compare the proposed model with other similar techniques in order to evaluate the AUC criterion in different schemes using ROC analysis. The movie genre was determined by comparing the model with GRU and ensGRU

spatial patterns in videos. GRU, as a recurrent network, has the ability to learn temporal patterns, while CNN is used to extract spatial features from video frames. This combination of two networks enables the model to identify more complex and diverse patterns, resulting in higher accuracy in video genre classification. One of the strengths of the ensGRU + CNN model is its ability to leverage the deep learning capabilities of CNN models while also being able to learn complex temporal patterns using GRU. This combination enhances the accuracy in detecting video genres. Additionally, the ensGRU + CNN model benefits from using pre-trained models such as ImageNet to extract useful features from video frames, improving the model's generalization ability. However, one potential weakness of this approach may be its high computational cost, especially when the model is dealing with large and complex datasets. Additionally, training this model may require access to large and suitable datasets, which may be challenging for some domains. When comparing the ensGRU + CNN approach with other methods such as VGG, ResNet, and DenseNet, distinct strengths and weaknesses emerge. VGG, known for its straightforward architecture and spatial feature extraction prowess, tends to suffer from computational inefficiency due to its parameter-heavy nature, and may struggle with capturing temporal patterns effectively. Similarly, while ResNet addresses the vanishing gradient problem and achieves excellent performance in image classification, its reliance on purely convolutional layers limits its ability to model temporal dependencies in video data. DenseNet's dense connections promote feature reuse and gradient flow, yet scalability issues persist as the network deepens, potentially hindering its performance in tasks requiring explicit modeling of temporal dynamics. In Fig. 6, a comparison has been made between the methods based on accuracy criteria, indicating that the proposed method exhibits superior genre classification capabilities in relation to video films.

Given the recent methods published by researchers in recent years, the superiority of the ensGRU + CNN model over them can be explained as follows:

1. Superiority over GRU + SVM [42]:
   - The ensGRU+CNN model utilizes a combination of GRU and CNN, providing a significant improvement in capturing both temporal and spatial patterns. In contrast, GRU+SVM relies solely on recurrent neural networks and a traditional classification method like SVM, which may offer less assistance in embedding spatial features for video classification.
2. Superiority over 1D-Conv-V [42]:
   - Compared to 1D-Conv-V, which employs a one-dimensional convolutional network for video analysis, the ensGRU+CNN model has a better ability to learn complex temporal and spatial features. This

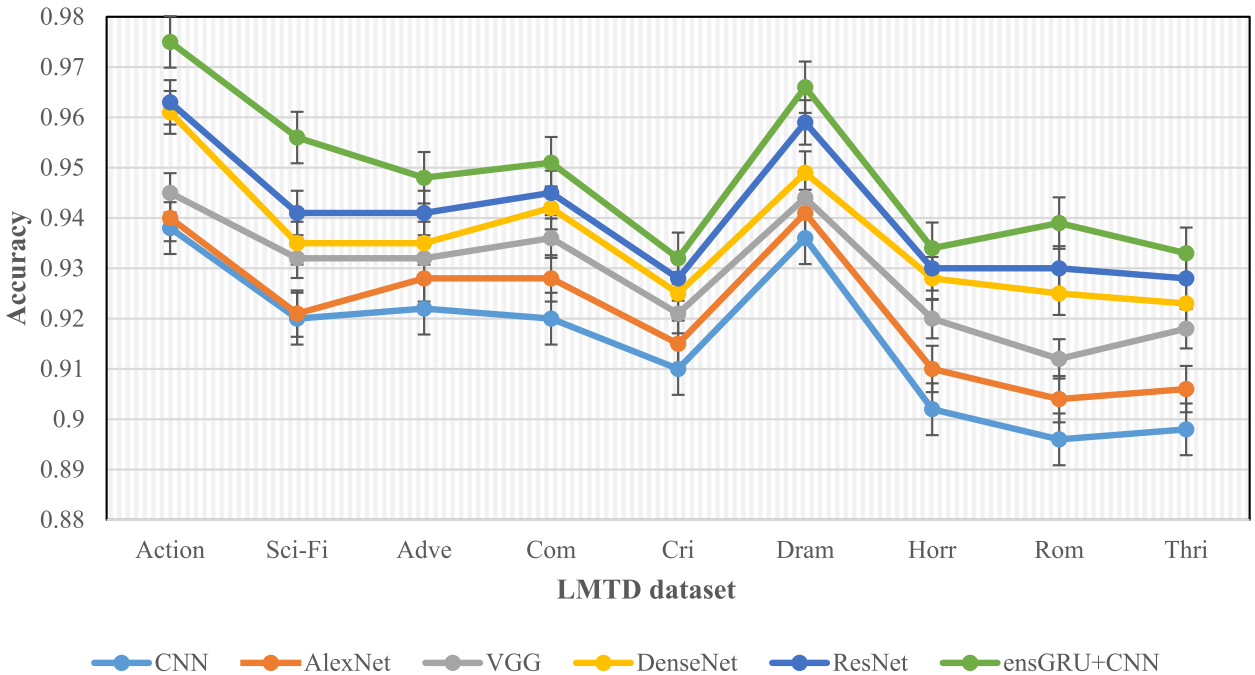

Fig. 6 In this figure, we compared our model with similar deep strategies based on accuracy criteria, demonstrating that the proposed method exhibits superior genre classification capabilities in relation to video films

fusion of GRU and CNN can enhance the recognition of intricate temporal patterns and increase classification accuracy.

3. Superiority over LLFM [37], LSTM [57], and CTT-MMC-TN [57]:
   - The Ensemble GRU+CNN model utilizes an effective combination of recurrent and convolutional neural networks to improve video classification accuracy. In contrast, LLFM, LSTM, and CTT-MMC-TN employ alternative methods for modeling and classifying, which may not be as effective in capturing temporal and spatial patterns.

Overall, the ensGRU + CNN model, with its blend of recurrent and convolutional neural networks, demonstrates better capabilities in modeling videos and recognizing complex temporal and spatial patterns, resulting in higher accuracy and precision in video genre classification. The results presented in Table 2 demonstrate a significant improvement in scores across various genres with the incorporation of the multi-modal component within the network. This observation implies that integrating movie soundtracks can greatly enhance the accuracy of genre prediction.

The observed decrease in the science fiction score with the inclusion of audio can be attributed to the inherent symmetry found in the musical compositions of action and science fiction films. Analysis of confusion matrices and loss convergence indicates a noticeable difference in the quantity of action movie trailers compared to sci-fi trailers. This discrepancy has the potential to create confusion within the network, resulting in misclassification of these two genres. The AUC ratings of the Gated Recurrent Unit (GRU) + SVM and 1D-Conv + SVM models show close proximity. However, it is evident that the GRU + SVM model outperforms the 1D-Conv + SVM model in five out of nine genres. Furthermore, the GRU + SVM technique demonstrates superior performance in all categories except for drama, as assessed by the AUC metric, when compared to the most advanced models. Notably, the GRU + SVM model shows significant improvements in performance within genres characterized by limited data availability, such as thriller and horror videos.

Limitations and difficulties

One of the primary limitations encountered in genre classification of videos using the Ensemble GRU + CNN method is data imbalance among different genres. Certain genres may have a significantly larger number of samples compared to others, leading to biased model training and potentially affecting the accuracy of classification for underrepresented genres. Another challenge lies in the complexity of temporal patterns present in video data. While the Ensemble GRU + CNN approach excels in capturing both spatial and temporal information, the intricate nature of temporal dynamics in videos, such as rapid scene changes or subtle transitions, can pose difficulties in accurate genre classification. Moreover, the subjective nature of genre definitions can introduce variability in the classification process. Different annotators may have diverse interpretations of genre labels, leading to inconsistencies in the labeled dataset. This variability can affect the model's ability to generalize across genres and may result in misclassification errors. Despite the effectiveness of deep learning models like Ensemble GRU + CNN, there may be limitations in their ability to understand contextual nuances within videos. Some genres may rely heavily on subtle cues or contextual information that is challenging for the model to discern accurately, potentially leading to misclassification or ambiguity in genre assignment. In addition, implementing the Ensemble GRU + CNN method for video genre classification requires significant computational resources, including powerful hardware for training and inference. The computational complexity of deep learning architectures, coupled with the need for large-scale video datasets, can pose challenges in terms of infrastructure and resource availability for researchers and practitioners.

Table 2 The table below provides a comprehensive overview of technical breakthroughs and the corresponding fields in which each film aims to implement its methodologies, focusing on accuracy criteria

Strategy	Accuracy								
	Action	Adventure	Comedy	Crime	Drama	Horror	Romance	Sci-Fi	Thriller
GRU + SVM [42]	0.909	0.929	0.910	0.846	0.829	0.865	0.725	0.628	0.825
1D-Conv-V [42]	0.777	0.807	0.803	0.659	0.733	0.810	0.655	0.787	0.763
LLFM [37]	0.852	0.752	0.871	0.628	0.641	0.424	0.468	0.192	0.520
LSTM [57]	0.687	0.573	0.792	0.421	0.740	0.478	0.313	0.237	0.437
CTT-MMC-TN [57]	0.835	0.672	0.870	0.547	0.841	0.667	0.456	0.401	0.522
CNN (VGG) + ensGRU	**0.974**	**0.948**	**0.952**	**0.932**	**0.968**	**0.935**	**0.941**	**0.956**	**0.935**

The decision to employ the proposed method over alternative approaches stems from several advantages that our technique offers, including enhanced performance, efficiency, and applicability to the specific problem domain of movie genre classification. Our method, which integrates a novel ensemble GRU-like unit, has demonstrated superior performance compared to existing techniques on benchmark datasets. Through extensive testing, we have consistently observed higher accuracy rates and F1 scores, indicating the effectiveness of our approach in accurately classifying video genres. Our ensemble GRU architecture is designed to efficiently capture spatial and motion data while revealing their temporal relationships. By leveraging the cross-correlation operator and employing convolution and cross-correlation techniques, our model can effectively extract relevant features from video frames, leading to improved classification accuracy. Furthermore, our method exhibits a high degree of applicability to the problem domain of movie genre classification. We have conducted thorough evaluations on well-known video genre datasets and have also validated the effectiveness of our approach in real-world scenarios, such as video production and maintenance, using a compiled assembly dataset. These experiments demonstrate the robustness and versatility of our proposed method across different domains and applications. In summary, the main reason for choosing our proposed method lies in its ability to deliver superior performance, efficiency, and applicability to the specific problem domain of movie genre classification.

Utilizing a single dataset allowed us to thoroughly evaluate the performance of our proposed method under controlled conditions. Despite the constraints, we have conducted extensive testing and validation on the LMDT database to provide a comprehensive assessment of our approach's effectiveness. Although we recognize the importance of testing on multiple datasets to validate the generalizability of our method, we believe that our results on the LMDT dataset provide valuable insights into the capabilities of our proposed model. Furthermore, fine-tuning across multiple datasets introduces additional challenges, such as dataset-specific biases and variations, which may require further adjustments to the model architecture and hyperparameters. Given these complexities, we opted to focus on thoroughly evaluating our proposed model on the LMDT dataset, ensuring a comprehensive understanding of its performance within a controlled setting.

Conclusion

Our study achieves a commendable classification accuracy of 94.4%. The proposed method, Ensemble GRU + CNN, emerges as an effective approach for online video genre recognition, leveraging the combined strengths of gated recurrent units (GRU) and convolutional neural networks (CNN). Through comparisons with several techniques in video genre classification, we justify the superior performance of our model. Additionally, we employed edge processing in our work to facilitate real-time processing and analysis of video data. By leveraging edge processing capabilities, we ensured smooth processing and relied on it to enhance the analysis of video data. This enabled us to achieve accurate classification across various genres while maintaining efficiency in digital media service management systems. The high classification accuracy attained underscores the efficacy of the Ensemble GRU + CNN model in digital media service management, offering promising prospects for enhanced content recommendation and personalized user experiences on online video platforms. As authors, we acknowledge several challenges and anticipate future research endeavors. Acquiring diverse and comprehensive datasets for video genre classification remains a significant challenge, necessitating labeled data spanning various genres and cultural backgrounds for robust model training. Achieving model generalization across different platforms and user preferences also poses challenges, requiring thorough experimentation and fine-tuning of the Ensemble GRU + CNN model. Implementing real-time inference of genre classification models in digital media service management systems introduces computational challenges, demanding optimization strategies tailored to platform-specific requirements. Incorporating user feedback and preferences into the genre classification pipeline adds complexity, necessitating the development of adaptive algorithms capable of dynamically adjusting genre classification based on user interactions and feedback. By acknowledging these challenges and incorporating edge processing into our approach, we aim to advance the field of online video genre recognition and digital media service management, paving the way for more efficient and accurate video analysis in real-time applications.

Acknowledgements

This work has received funding from the 2022 Innovative Application of Virtual Simulation Technology in Vocational Education Teaching Special Project of the Higher Education Science Research and Development Center of the Ministry of Education - Research on the Construction and Application of Virtual Simulation Dance Teaching Full Scene Based on Action Capture Technology (ZJXF2022177).

Authors' contributions

Y. S. wrote the main manuscript text; N. G. reviewed the manuscript.

Funding

This work has received funding from the 2022 Innovative Application of Virtual Simulation Technology in Vocational Education Teaching Special Project of the Higher Education Science Research and Development Center of the Ministry of Education—Research on the Construction and Application of Virtual Simulation Dance Teaching Full Scene Based on Action Capture Technology (ZJXF2022177).

Availability of data and materials
Data is provided within the manuscript or supplementary information files. https://github.com/jwehrmann/lmtd. https://dl.acm.org/doi/abs/10.1145/3019612.3019641.

Declarations

Competing interests
The authors declare no competing interests.

Received: 20 March 2024 Accepted: 3 May 2024
Published online: 14 May 2024

References

1. Chen Z, Ye S, Chu X, Xia H, Zhang H, Qu H, Wu Y (2021) Augmenting sports videos with viscommentator. IEEE Trans Visual Comput Graphics 28(1):824–834
2. Almeida A, de Villiers JP, De Freitas A, Velayudan M (2022) The complementarity of a diverse range of deep learning features extracted from video content for video recommendation. Expert Syst Appl 192:116335
3. Mahadevkar SV, Khemani B, Patil S, Kotecha K, Vora DR, Abraham A, Gabralla LA (2022) A review on machine learning styles in computer vision—Techniques and future directions. IEEE Access 10:107293–107329
4. Rezaee K et al (2024) A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Pers Ubiquit Comput 28(1):135–151
5. Huang Q, Xiong Y, Rao A, Wang J, Lin D (2020) Movienet: A holistic dataset for movie understanding. Computer Vision–ECCV 2020: 16th European Conference. Springer, Glasgow, UK (August 23–28, 2020. Proceedings, Part IV, 709–727)
6. Huang Q, Xiong Y, Xiong Y, Zhang Y, & Lin D. (2018). From trailers to storylines: An efficient way to learn from movies. arXiv preprint arXiv:1806.05341
7. Deldjoo Y, Elahi M, Cremonesi P, Garzotto F, Piazzolla P, Quadrana M (2016) Content-based video recommendation system based on stylistic visual features. Journal on Data Semantics 5:99–113
8. Lee J, Abu-El-Haija S (2017) Large-scale content-only video recommendation. Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops
9. Dastbaravardeh E, et al (2024) Channel Attention-Based Approach with Autoencoder Network for Human Action Recognition in Low-Resolution Frames. Int J Intell Syst 2024:1–22. Article ID: 1052344. https://doi.org/10.1155/2024/1052344
10. Montalvo-Lezama R, Montalvo-Lezama B, Fuentes-Pineda G (2023) Improving Transfer Learning for Movie Trailer Genre Classification using a Dual Image and Video Transformer. Inf Process Manage 60(3):103343
11. Abu-El-Haija S, Kothari N, Lee J, Natsev P, Toderici G, Varadarajan B, Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675
12. Brezeale D, & Cook DJ. (2006). Using closed captions and visual features to classify movies by genre. In Poster session of the seventh international workshop on Multimedia Data Mining (MDM/KDD2006). Citeseer, MDM/KDD'06, Philadelphia
13. Rezaee K, et al, (2023). IoMT-assisted medical vehicle routing based on UAV-Borne human crowd sensing and deep learning in smart cities. IEEE Internet of Things J 10(21):18529–18536. https://doi.org/10.1109/JIOT.2023.3284056
14. Fenercioglu L, Türköz I, Güvenir A (2022) Movie Trailer Scene Classification Based on Audio VGGish Features. 2022 International Conference on Machine Learning, Control, and Robotics (MLCR). pp 49–54
15. Rajput NK, Grover BA (2022) A multi-label movie genre classification scheme based on the movie's subtitles. Multimedia Tools and Applications 81(22):32469–32490
16. Zhou B, Andonian A, Oliva A, Torralba A (2018) Temporal relational reasoning in videos. Proceedings of the European Conference on Computer Vision (ECCV). pp 803–818
17. Feichtenhofer C, Fan H, Malik J, He K (2019) Slowfast networks for video recognition. Proceedings of the IEEE international conference on computer vision. pp 6202–6211
18. Xiao F, Lee YJ, Grauman K, Malik J, & Feichtenhofer C. (2020). Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740
19. Zhang Z, Gu Y, Plummer BA, Miao X, Liu J, Wang H (2024) Movie genre classification by language augmentation and shot sampling. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2024:7275–7285
20. Unal FZ, Guzel MS, Bostanci E, Acici K, Asuroglu T (2023) Multilabel Genre Prediction Using Deep-Learning Frameworks. Appl Sci 13(15):8665
21. Cai Z, Ding H, Wu J, Xi Y, Wu X, Cui X (2023) Multi-label movie genre classification based on multimodal fusion. Multimedia Tools and Applications 13(15):8665
22. Cascante-Bonilla, P., Sitaraman, K., Luo, M., & Ordonez, V. (2019). Moviescope: Large-scale analysis of movies using multiple modalities. arXiv preprint arXiv:1908.03180
23. Yang X, Zhou Q, Chen W, Zhao L (2023) MFMGC: A Multi-modal Data Fusion Model for Movie Genre Classification. International Conference on Advanced Data Mining and Applications. Cham Springer, Nature Switzerland, pp 676–691
24. Bain M, Nagrani A, Brown A, Zisserman A (2020) Condensed movies: Story based retrieval with contextual embeddings. Proceedings of the Asian Conference on Computer Vision
25. Bi T, Jarnikov D, Lukkien J (2022) Shot-Based Hybrid Fusion for Movie Genre Classification. International Conference on Image Analysis and Processing 2022. Springer International Publishing, Cham, pp 257–269
26. Pant P, Sai Sabitha A, Choudhury T, Dhingra P (2019) Multi-label classification trending challenges and approaches. In Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018:433–444
27. Oh J, Sudarshan S, Lee JA, Yu N (2022) Serendipity enhances user engagement and sociality perception: The combinatory effect of serendipitous movie suggestions and user motivations. Behaviour & Information Technology 41(11):2324–2341
28. Sim G. (2023). The Idea of Genre in the Algorithmic Cinema. Television & New Media, 15274764231171072
29. Rasheed Z, Sheikh Y, Shah M (2005) On the use of computable features for film classification. IEEE Trans Circuits Syst Video Technol 15(1):52–64
30. Jain SK, Jadon RS (2009) Movies genres classifier using neural network. 2009 24th International Symposium on Computer and Information Sciences. pp 575–580
31. Huang, Y. F., & Wang, S. H. (2012). Movie genre classification using SVM with audio and video features. In Active Media Technology: 8th International Conference, AMT 2012, Macau, China, December 4–7 (2012) Proceedings 8 2012. Springer, Berlin Heidelberg, pp 1–10
32. Zhou H, Hermans T, Karandikar AV, Rehg JM (2010) Movie genre classification via scene categorization. Proceedings of the 18th ACM international conference on Multimedia 2010. pp 747–750
33. Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. Int J Comput Vision 42:145–175
34. Wu J, Rehg JM (2008) Where am I: Place instance and category recognition using spatial PACT. 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp 1–8
35. Simoes GS, Wehrmann J, Barros RC, Ruiz DD (2016) Movie genre classification with convolutional neural networks. 2016 International Joint Conference on Neural Networks (IJCNN). pp 259–266
36. Ogawa T, Sasaka Y, Maeda K, Haseyama M (2018) Favorite video classification based on multimodal bidirectional LSTM. IEEE Access 6:61401–61409
37. Álvarez F, Sánchez F, Hernández-Peñaloza G, Jiménez D, Menéndez JM, Cisneros G (2019) On the influence of low-level visual features in film classification. PLoS ONE 14(2):e0211406
38. Ben-Ahmed O, Huet B (2018) Deep multimodal features for movie genre and interestingness prediction. 2018 international conference on content-based multimedia indexing (CBMI) IEEE. pp 1–6
39. Aytar Y, Vondrick C, & Torralba A. (2016). Soundnet: Learning sound representations from unlabeled video. Adv Neural Inf Process Syst 29:1–9.

https://proceedings.neurips.cc/paper/2016/file/7dcd340d84f762eba80aa538b0c527f7-Paper.pdf
40. Yu Y, Lu Z, Li Y, Liu D (2021) ASTS: attention based spatio-temporal sequential framework for movie trailer genre classification. Multimedia Tools and Applications 80:9749–9764
41. Varghese J, Ramachandran Nair KN (2019) A novel video genre classification algorithm by keyframe relevance. Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, vol Volume 1. Springer, Singapore, pp 685–696
42. Choroś K (2019) Fast method of video genre categorization for temporally aggregated broadcast videos. Journal of intelligent & fuzzy systems 37(6):7657–7667
43. Yadav A, Vishwakarma DK (2020) A unified framework of deep networks for genre classification using movie trailer. Appl Soft Comput 96:106624
44. Jiang Y, Zheng L (2023) Deep learning for video game genre classification. Multimedia Tools and Applications. pp 1–5
45. Mangolin RB, Pereira RM, Britto AS Jr, Silla CN Jr, Feltrim VD, Bertolini D, Costa YM (2022) A multimodal approach for multi-label movie genre classification. Multimedia Tools and Applications 81(14):19071–19096
46. Behrouzi T, Toosi R, Akhaee MA (2023) Multimodal movie genre classification using recurrent neural network. Multimedia Tools and Applications 82(4):5763–5784
47. Chun-Fu RC, Rameswar P, Kandan R, Rogerio F, John C, Aude O, Quanfu F (2021) Deep analysis of cnn-based spatio-temporal representations for action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 6165–6175
48. Yue M, Chung-Ching L, Rameswar P, Prasanna S, Leonid K, Aude O, Kate S, Rogerio F (2020) Ar-net: Adaptive frame resolution for efficient action recognition. European Conference on Computer Vision. 86, 104
49. Du T, Lubomir B, Rob F, Lorenzo T, Manohar P (2015) Learning spatiotemporal features with 3d convolutional networks. Proceedings of the IEEE international conference on computer vision. pp 4489–4497
50. Limin W, Yuanjun X, Zhe W, Yu Q, Dahua L, Xiaoou T, Van Luc G (2016) Temporal segment networks: Towards good practices for deep action recognition. European conference on computer vision. Springer, pp 20–36
51. Ba TT, Svetha V (2007) Video abstraction: A systematic review and classification. ACM transactions on multimedia computing, communications, and applications (TOMM) 3(1):3-es
52. Danila P, Matthijs D, Zaid H, Cordelia S (2014) Category-specific video summarization. European conference on computer vision. Springer, pp 540–555
53. Wei S et al (2023) Edge-enabled federated sequential recommendation with knowledge-aware Transformer. Futur Gener Comput Syst 148:610–622
54. Rocha Neto A, Silva TP, Batista T, Delicato FC, Pires PF, Lopes F (2020) Leveraging edge intelligence for video analytics in smart city applications. Information 12(1):14
55. Tang, S., et al., (2023). Edge Intelligence with Distributed Processing of DNNs: A Survey. CMES-Computer Modeling in Engineering & Sciences 136(1). https://doi.org/10.32604/cmes.2023.023684.
56. Wehrmann J, Barros RC (2017) Convolutions through time for multi-label movie genre classification. Proceedings of the Symposium on Applied Computing 2017 Apr 3. pp 114–119
57. Wehrmann J, Barros RC (2017) Movie genre classification: A multi-label approach based on convolutions through time. Appl Soft Comput 61:973–982

Publisher's Note