Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Movie Scene Genre Classification & Video Recommendation System using Deep Learning

Ahsan Arshad
Institute of Management Sciences, Peshawar. ahsanarshad777@gmail.com
Afsheen Khalid
Institute of Management Sciences, Peshawar. afsheen.khalid@imscineces.edu.pk
Dilawar Khan
University of Engineering and Technology, Peshawar.
dilawarkhan@uetpeshawar.edu.pk
Sumaira Johar
Institute of Management Sciences, Peshawar. sumaira.johar@imsciences.edu.pk
Omar bin Samin
Institute of Management Sciences, Peshawar. omar.samin@imsciences.edu.pk

ABSTRACT

Object and scene classification methods have often struggled to match human performance due to background complexity, context understanding and complex foreground elements that often appear blurry or moving. These are particularly evident in movie analysis with dynamic backgrounds and frames containing complex features. With the recent advancement and development of deep learning and convolutional neural networks (CNN), numerous possibilities have open for better scene categorization and higher-level tasks such as genre classification. This study aims to build up on the ---- and proposes a transfer learning based CNN model that utilizes inception V3 model. the study utilized 80 videos from YouTube that contained movie scenes over four genre types; Action, Horror, Sci-Fi and Romance. The extracted scenes from each video were analyzed and Euclidean Distance method was applied for measuring dissimilarity and to remove any redundant scenes. The model was cross tested on unseen data and evaluated using Precision, Recall, Accuracy and F1 score. The base line model achieved a 66% accuracy which was further improved by fine-tuning the mode to 72%. Furthermore, the model achieved a 67% Precision, 64% Recall and a 65% F1 score. The result demonstrated that the proposed transfer learned model showed accuracy surpassing human classification abilities. The findings of the research can further aid in the applications related to

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

content tagging, recommendation systems, film analytics and copyright monitoring, by providing a faster and scalable genre classification system.

Keywords: Deep Neural Network, Transfer Learning, Movie Recommendation, Scenes Detection

INTRODUCTION

Each year, a large number of films and TV shows are released, and the overall quantity of movie and TV titles is astonishing. 3.25 million titles are currently accessible through IMDB [1], one of the biggest databases for films and television. Another truth is that, over the past four years, more people have watched films and TV episodes online than ever before, with monthly time spent viewing Netflix content increasing from 28.3 hours in 2011 to 46.6 hours in 2014 [1]. Making better recommendations from the enormous selection of films and TV shows is the fundamental issue in the scenario. The common recommendation techniques, such as collaborative filtering, leverage other users' viewing patterns and ratings to enhance your suggestion. As a result, it frequently suggests the most well-liked films, which makes personalized recommendation more difficult. The information taken from the actual movie material, however, may be a great addition to the suggestion because it is not impacted by user activity. Scene identification is very important since the presence of a certain scene type may tell you a lot about a film.

In addition to the growing demand for video content extraction, video search is made possible by recent developments in computer vision, namely in the areas of information retrieval and deep neural networks. Recent studies on identifying images used Deep Neural Networks, with remarkable results in terms of both accuracy and the number of detectable classes. There are now over 14 million images and 21841 classes in ImageNet [1], the benchmark dataset for object classification. Regarding the information retrieval component of the issue, an increasing number of businesses, such as Google and Microsoft, are attempting to transition from text-based image retrieval to content-based image retrieval. The key element of this shift is information extraction from the image itself. Scene information is undoubtedly one of the bits of data extracted from both videos and photos. The capacity to process any film and extract scene or environment information would also generally give viewers with more tailored adverts, therefore the scope of this study paper is

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

not limited to only movie content analysis.

Inception-V3 is an image recognition and a convolutional neural network model trained on more than hundred thousand images of the dataset of the Image-Net, a benchmark dataset having 1000 different classes, it is a 315-layer pre-trained neural network model composed of two parts, featuring extraction layers from 1 to 313 with a convolutional operation and last three i.e., 314-315 are classification layers which are fully connected followed by SoftMax and classification layer.

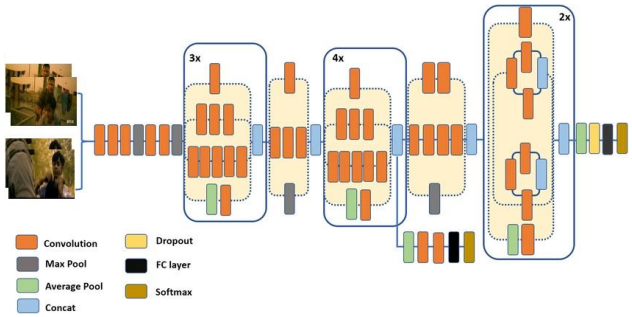


Figure 1 An Illustration of InceptionV3 Architecture [18]

Despite the availability of a vast amount of movie scene frame data and the application of machine learning (ML) techniques, achieving a high level of performance in movie scene analysis remains a challenge. ML algorithms, including Convolutional Neural Networks (CNNs), often struggle to effectively extract hidden patterns (features) from movie scene frames. These hidden patterns play a crucial role in tasks such as object recognition, scene classification, and genre categorization.

One key issue is the inability of ML algorithms to handle non-invariant features, such as variations in scale, rotation, illumination, and translation within movie scenes. Movie frames frequently exhibit complex visual transformations, making it difficult for traditional ML approaches to generalize and recognize objects or scenes across different instances. Moreover, ML algorithms' limited capability to extract non-invariant features hinders their ability to accurately analyze and categorize movie scenes. Hidden patterns that convey important semantic information may be overlooked or improperly represented, leading to inaccurate scene recognition and genre classification.

The challenge of extracting non-invariant features becomes even more pronounced due to the large-scale nature of movie scene datasets. The sheer volume of data poses computational challenges and requires efficient and scalable ML

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

algorithms capable of handling the intricacies of movie scenes while maintaining high performance.

Therefore, the primary objective of this research is to address the limitations of ML algorithms in extracting non-invariant features from movie scene frames. The research aims to investigate and propose novel approaches to overcome these limitations, taking into account the challenges posed by scale, rotation, illumination, and translation variations. The following key questions will guide the research:

1) What are the specific limitations of ML algorithms, including CNNs, in extracting non-invariant features from movie scene frames?
2) How do variations in scale, rotation, illumination, and translation impact the accurate extraction and representation of hidden patterns within movie scenes?
3) What are the implications of ML's limited ability to handle non-invariant features for tasks such as object recognition, scene classification, and genre categorization?
4) What innovative techniques and algorithms can be developed to address the challenges associated with non-invariant feature extraction from movie scene frames?
5) How can the proposed techniques be evaluated and benchmarked against existing methods to demonstrate their effectiveness in improving scene recognition and genre categorization accuracy?

By addressing these research questions, this study aims to advance the field of ML and computer vision by proposing novel solutions for extracting non-invariant features from movie scene frames. The outcomes of this research will contribute to the development of more robust and accurate ML algorithms for movie analysis, genre categorization, and scene recognition, overcoming the limitations posed by non-invariant features.

Research on developing machine learning methods for image classification for one-label classification is currently in process. Nevertheless, it has been demonstrated that video-based methods are more difficult and have not yet been fully researched. Recent studies have shown that convolutional neural networks (CNN) can produce better outcomes in this field.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

The objective of this study is to use the CNN methods to one label and multi label movie genre classification because convolutional neural networks may avoid human feature extraction and because they perform well in object identification. It should be noted that the primary objective of this project is to comprehend the semantics of the video, and the classification of movie genres is an ideal place to start. In order to comprehend ongoing events and respond appropriately and promptly, future research can build on these findings of semantic comprehension on a variety of films, such as those taken by CCTV cameras or traffic cameras.

This research work proposes a method for training a movie scene detection model using deep learning techniques. Instead of relying on hand-crafted features, the proposed approach leverages the power of the Inception-v3 Convolutional Neural Network (CNN) model for feature extraction and transfer learning for classification. The model will be trained to classify movie scenes into four genres: action, horror, romantic, and sci-fi.

The proposed solution involves the following steps:

1) **Dataset Preparation:** A dataset comprising movie scenes from various genres will be collected and labeled. Each video will be processed to extract frames at a fixed rate of 30 frames per second (FPS). The dataset will be divided into a training set (80%) and a testing set (20%) using cross-validation.
2) **Feature Extraction:** The frames from the videos will be resized to the required input size of the Inception-v3 CNN model. Deep CNN features will be extracted from each frame using the pre-trained Inception-v3 model. This step captures high-level visual representations from the frames, which will be used for classification.
3) **Model Training:** The extracted features will be used to train a classification model using transfer learning. The model will be trained on the training set, learning to differentiate between the four genres. Evaluation metrics such as accuracy, recall, precision, and f-measure will be used to assess the performance of the trained model.
4) **Performance Evaluation:** The trained model will be tested on the testing set to evaluate its performance. The evaluation metrics will be calculated to

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

measure the accuracy, recall, precision, and f-measure of the model. The results will be compared with state-of-the-art systems to assess the effectiveness of the proposed approach.

5) **Development of a GUI-Based Application:** A simulation-based GUI application will be developed, incorporating deep learning and computer vision technologies. This application will accept a video as input and perform an analysis. The frames of the video will be classified using the trained convolutional neural network, providing details such as the number of actions in the video and the classification of each frame. The recommended class will be determined based on the genre with the maximum number of frames. The application will also visualize the distribution of frames using bar charts and pie charts. Additionally, metadata about the video, including its name, resolution, type, file format, duration, and FPS, will be displayed.

The proposed solution aims to enhance movie scene detection by leveraging deep learning and computer vision techniques. Through the use of transfer learning with the Inception-v3 model, accurate genre classification can be achieved. The GUI-based application will provide a user-friendly interface for analyzing videos and visualizing the scene distribution, offering insights into the content and characteristics of the input videos.

For scene recognition, this project uses the currently popular Convolutional Neural Network technique. I have generated a unique movie scene dataset and movie frames with scene labels. With at least 3600 training pictures for each of the four categories in the Custom Movie Scene dataset, an autonomous video analysis and recommendation system is being developed through the use of Transfer Learning in certain pretrain CNN models. Individual picture testing is used to evaluate the categorization of scenes. use the accuracy and F1-Score assessment measures to conduct a performance study of the trained models.

LITERATURE REVIEW

Artificial Neural Networks (ANNs), which are a type of machine learning model that is loosely based on the biology of the human brain and how it functions. They are employed in the learning process of recognizing patterns in data. The term "neural networks" is also used in some circles to refer to these systems. Unlike previous

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

generations of models, this generation is able to "learn" from examples that are provided to them and perform specific tasks that were previously impossible for them to complete successfully.

These networks are comprised of a collection of units or nodes that are connected to one another in a manner that is similar to the neural architecture of the human brain at the cellular level in general. In this system the artificial neurons that comprise it serve as the structural elements of the system. In the absence of genetic mutations, it is still possible to "learn" how strong the connections between pairs of neurons are by employing a process known as "model training," which is a type of cognitive training. A technique known as backpropagation is used in this process [2].

Backpropagation is a technique that allows for the adjustment of individual connection values (also known as weights) in order to minimize a function that represents (hopefully) how large the errors are that the network commits while performing the task at hand. The neural network from the machine learning and deep learning are widely used in the detection of various human action that can normal or anomaly based. As violence detection is challenging task and need lot of data for model training, so in that case the deep learning model performs well in systems such as a violence detection system.

In the research article [3], the author developed a video recommendation system using the seq2seq LSTM model. The proposed model is validated using the MovieLens dataset. The experimental results achieved by the RNN-LSTM is 96.27% on the training set while the accuracy achieved on the validation set is 95.89%.

In the research article [4], the author proposed an SVM based recommender system. The performance of the proposed system is enhanced using the best features section using Genetic Algorithm and hyper-parameters tunning. The accuracy achieved by the SVM model is 66% while the GASVM has outperformed SVM with an accuracy of 67%.

In research paper [10] the scene classification and recognition have long been fundamental challenges in computer vision, with significant implications for applications ranging from content-based image retrieval to autonomous systems. As

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

reviewed by Singh et al. (2017), the field has evolved through two primary approaches: object-based recognition (SR-OR) and low-level feature-based methods (SR-LLF). Object-based techniques, exemplified by the Object Bank framework (Li et al., 2010), leverage object detectors to infer scene semantics but face challenges with computational complexity and error propagation. In contrast, low-level feature methods employ texture analysis (Renninger & Malik, 2004), orientation features (Gorkani & Picard, 1994), and bag-of-features models (Csurka et al., 2004) to achieve robust performance despite occlusion and lighting variations. More recent probabilistic approaches, including Latent Dirichlet Allocation (Fei-Fei & Perona, 2005) and deformable part models (Pandey & Lazebnik, 2011), have bridged these paradigms by incorporating spatial relationships and semantic context. The emergence of deep learning has further transformed the field, with convolutional neural networks demonstrating superior performance in learning hierarchical scene representations. However, challenges remain in handling scene ambiguity, scaling to diverse real-world environments, and achieving human-like holistic understanding. Future directions likely involve the integration of multimodal cues, explainable AI techniques, and continued advances in transfer learning to address these limitations while building on the rich foundation of traditional computer vision methods.

The paper [11] by Bose et al. addresses the challenge of visual scene recognition in long-form media, particularly movies, which present unique complexities due to their narrative structures, diverse visual settings, and frequent scene transitions. The authors highlight key limitations in existing datasets, such as the lack of comprehensive taxonomies tailored to movie scenes and the absence of shot-level annotations, which are crucial for capturing fine-grained scene transitions. Previous works in scene recognition have primarily focused on natural images (e.g., MIT Indoor67, SUN397, Places) or short web videos (e.g., HVU, SOA), often neglecting the domain-specific challenges posed by movies, such as actor-centric framing, fictional settings, and dynamic camera movements. Movie-centric datasets like Hollywood Scenes, Moviegraphs, and Movienet have made strides but remain limited in scope, either covering a small number of scene classes or lacking detailed shot-level annotations.

To bridge this gap, the authors propose a novel movie-centric taxonomy of 179

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

scene labels, automatically curated using scene headers (sluglines) from movie scripts and auxiliary web-video datasets like HVU, with minimal human intervention. This taxonomy captures a broader range of real-world and fictional settings compared to prior works. Instead of relying on costly manual annotations, the paper leverages CLIP, a vision-language model, to weakly label 1.12 million movie shots from the Condensed Movies dataset. This weakly labeled dataset, called MovieCLIP, enables the training of baseline models for multi-label scene classification. The authors evaluate these models on a human-verified subset and demonstrate their effectiveness in downstream tasks such as web-video scene classification (HVU) and movie genre prediction (Moviescope). Their approach shows that pretraining on MovieCLIP improves performance over Kinetics-pretrained models, particularly in movie-related tasks, highlighting the value of domain-specific datasets.

The paper situates itself within broader trends in vision-language pretraining, where models like CLIP are increasingly used for zero-shot and weakly supervised learning. Unlike prior works that rely on manual annotations or crowd-sourced labels, this study demonstrates how large-scale pretrained models can automate dataset curation while maintaining reasonable accuracy. The ethical implications of such automated labeling are also discussed, emphasizing its potential to analyze biases in media representation. Overall, this work advances scene recognition in movies by introducing a scalable annotation framework, a rich taxonomy, and models that generalize to related tasks, setting a foundation for future research in multimodal movie understanding.

The paper [12] by Deldjoo et al. presents a novel approach to video recommendation by leveraging low-level stylistic visual features—such as lighting, color variance, and motion extracted directly from video content. Unlike traditional content-based recommender systems (CBRS) that rely on high-level metadata (e.g., genre, cast, or director), this work explores the potential of implicit visual cues to improve recommendation accuracy, particularly in scenarios where metadata is sparse or unavailable (e.g., user-generated content on platforms like YouTube). The authors ground their approach in *Applied Media Aesthetics*, a theoretical framework that links visual elements (e.g., lighting key, shot length) to emotional and

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

perceptual responses in viewers, arguing that these features can serve as robust proxies for user preferences.

Prior research in video recommendation has largely focused on collaborative filtering or hybrid methods combining metadata and user behavior, with limited exploration of low-level visual features. Some exceptions include works like *VideoReach* (Zhang et al., 2013) and multi-modal approaches (Xu et al., 2015), which integrate visual features alongside textual and aural data but do not isolate their impact. Deldjoo et al. differentiate their work by demonstrating that stylistic features alone can outperform genre-based recommenders, achieving higher recall in offline evaluations. Additionally, they validate that trailers despite being shorter exhibit strong feature correlation with full-length movies, making them viable substitutes for computationally expensive full-video processing.

The study employs a k-nearest neighbor (KNN) algorithm to generate recommendations, comparing performance across feature combinations. Key findings include: (1) recommendations based on lighting and motion features outperform genre-based methods by up to 50% in recall, (2) color variance is less discriminative, and (3) combining low- and high-level features does not improve accuracy, likely due to collinearity. The authors also highlight the entropy and correlation analyses to justify feature selection, noting that trailers preserve the stylistic essence of full movies.

This work contributes to the broader field of multimedia recommender systems by addressing the *"extreme new item problem"* where new videos lack metadata and by empirically validating the role of stylistic features in user preference modeling. Future directions include extending the feature set (e.g., incorporating audio), testing larger datasets, and evaluating perceived novelty and diversity in user studies. The paper bridges computational techniques with media theory, offering a scalable solution for platforms with sparse metadata while underscoring the artistic principles that shape viewer engagement.

The paper [13] by Khadka and Lamichhane presents a content-based video recommendation system that leverages textual metadata (genre, cast, and overview) to suggest relevant content to users. Unlike collaborative filtering, which relies on user behavior patterns, this approach focuses on item attributes, making it

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

particularly useful for new users or items with limited interaction data a common challenge known as the *cold-start problem*. The authors employ **TF-IDF (Term Frequency-Inverse Document Frequency)** for text vectorization, converting unstructured movie descriptions into numerical representations, followed by **cosine similarity** to measure content similarity and generate recommendations.

Prior research in recommendation systems has explored various methodologies, including collaborative filtering (e.g., Amazon's item-to-item approach) and hybrid models combining multiple techniques. However, content-based systems are particularly advantageous in scenarios where user interaction data is sparse. The authors build upon foundational work in **TF-IDF-based recommendations**, a well-established technique in information retrieval, but adapt it for video streaming by focusing on movie metadata rather than user ratings. Their approach aligns with earlier studies demonstrating that content-based filtering can effectively capture user preferences without requiring extensive historical data from multiple users.

The paper evaluates system performance using standard metrics **precision (90.9%), recall (83.3%), and F1-score (86.95%)** derived from a confusion matrix analysis. These results suggest that the model performs well in recommending relevant content, though the study is limited by a small sample dataset (five movies for initial testing). Future work could expand the dataset and incorporate **hybrid techniques**, blending content-based and collaborative filtering to further improve accuracy. Additionally, integrating **deep learning-based embeddings (e.g., Word2Vec, BERT)** could enhance semantic understanding beyond TF-IDF's keyword-based approach.

This research contributes to the growing field of personalized video recommendations by demonstrating the viability of a lightweight, metadata-driven system, particularly useful for platforms with limited user interaction data. The findings underscore the importance of feature selection (genre, cast, and plot) in content-based models while highlighting avenues for future enhancements, such as larger-scale validation and advanced NLP techniques.

The chapter [14] by Pazzani and Billsus provides a comprehensive overview of

Online ISSN 3007-3197
Print ISSN 3007-3189

content-based recommendation techniques, focusing on systems that leverage item descriptions and user profiles to generate personalized suggestions. The authors distinguish between structured data (e.g., databases with fixed attributes like cuisine or cost) and unstructured data (e.g., free-text descriptions), highlighting the challenges and methodologies for handling each type. Structured data often employs traditional machine learning algorithms like **decision trees** and **rule induction**, which work well with well-defined attributes, while unstructured text requires **vector space models** (e.g., TF-IDF) or **probabilistic methods** (e.g., Naïve Bayes) to convert text into a structured format for analysis.

The chapter reviews several key learning algorithms used in content-based recommendation systems, including **nearest neighbor methods**, **Rocchio's algorithm**, **linear classifiers**, and **Naïve Bayes**, each suited to different data representations and scalability needs. For instance, nearest neighbor methods rely on similarity metrics like cosine similarity, making them effective for text-based recommendations, whereas Naïve Bayes excels in probabilistic classification tasks. The authors also discuss **relevance feedback** mechanisms, where user interactions refine recommendations over time, enhancing personalization.

A notable limitation of content-based systems is their reliance on descriptive content if item descriptions lack discriminative features (e.g., jokes or abstract art), recommendations may falter. The authors suggest hybrid approaches combining **collaborative filtering** to mitigate this issue, as seen in systems like Amazon, which blend user preferences with item attributes. Additionally, the chapter touches on emerging trends, such as **support vector machines (SVMs)** for improved generalization in high-dimensional text classification tasks.

Overall, this work underscores the versatility of content-based recommendation systems across domains (e.g., news, e-commerce, entertainment) while acknowledging their constraints. Future directions include integrating richer semantic representations (e.g., word embeddings) and hybrid models to enhance recommendation accuracy and coverage. The chapter serves as a foundational reference for understanding the interplay between content representation, user modeling, and algorithmic choice in personalized recommendation systems.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

The paper [15] by Reddy et al. presents a content-based recommendation approach that leverages movie genres to personalize suggestions for users. Unlike collaborative filtering, which relies on user-user similarity, this system analyzes a user's past ratings to identify preferred genres and recommends movies with similar genre profiles. The study utilizes the MovieLens dataset, comprising 9,126 movies across 11 genres and ratings from 671 users, converting ratings into binary values (1 for ratings >3, −1 otherwise) and genres into binary vectors (1 if present, 0 otherwise). A dot product between the genre and rating matrices generates a user profile, and Euclidean distance measures identify the closest matches for recommendations.

The authors contextualize their work within the broader landscape of recommender systems, contrasting content-based, collaborative, and hybrid methods. Collaborative filtering, exemplified by platforms like Netflix and Amazon, suffers from the *cold-start problem* (new users/items lack sufficient interaction data), while content-based systems mitigate this by focusing on item attributes (e.g., genres). Hybrid systems, as noted, combine both approaches to balance their weaknesses e.g., Debnath et al.'s feature-weighting hybrid model and Wakil et al.'s emotion-based filtering. However, the paper argues that pure content-based systems, when optimized (e.g., via binary genre-weighting), can effectively personalize recommendations without requiring extensive user interaction data.

Key contributions include a streamlined algorithmic pipeline: (1) binary encoding of genres and ratings, (2) dot-product-based user profiling, and (3) Euclidean distance-based similarity matching. Results demonstrate genre-driven recommendations for sample users, validating the method's feasibility. Limitations include the simplicity of binary genre representation, which ignores nuanced genre interactions (e.g., a user preferring *sci-fi thrillers* over pure *sci-fi*). Future work could integrate richer metadata (e.g., directors, keywords) or hybridize with collaborative filtering for improved accuracy. The study underscores the commercial viability of content-based systems in scenarios where genre is a strong predictor of preference, aligning with industry applications in streaming platforms.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Movie Scene Recognition

A simple movie clip consists of a series of picture frames, the soundtrack, and the relevant subtitles. Therefore, compared to standard picture scene identification, we are given a lot more information for the movie scene recognition problem. As a consequence, we may examine frames that are part of the same scene and provide a more precise categorization result.

On the other hand, because movies shots are typically person-specific and feature close-ups of the primary figure, they can be more difficult to recognize than image scenes. To make things worse directors prefer to blur out the surroundings and concentrate on the performer to give a single shot an artistic vibe. Because the information we are trying to extract is all in the background, these features make it challenging to identify movie sequences. When the background is too unclear to be recognized or the character's face takes up too much of the frame, it is challenging to do an accurate categorization.

In summary, this article aims to minimize the limitations of a movie-specific dataset while maximizing the rich information provided by the movie clip.

Feature Extraction And Classification Methods

A popular method for solving image recognition issues is feature extraction, which provides a more accurate representation of an image than raw pixel values and this fed into machine learning algorithms to identify a certain class of objects. We shall review related research in this subject because different feature extraction methods will undoubtedly produce varied recognition results.

Convolutional Neural Network For Scene Recognition

Convolutional neural networks have been used in almost all image recognition successes in recent years. With the development of effective GPU computing, it is now feasible to train and implement enormous networks with even larger databases. It is so enormous that it can even replicate human perception and produce outcomes that are similar to those of human perception. Therefore, it becomes sense to take a quick glance at the history of CNN approaches here.[17]

Following the identification of visual systems in living things, the concept of a convolutional neural network was first proposed in the early 1980s. Each individual neuron in this type of feed forward artificial neural network reacts to an

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

overlapping area in the visual field before its own layer.

METHODOLOGY

We examine the research report project's methodology in detail in this chapter. In addition to providing information on how to create and train a convolutional neural network, it also explains how to establish a database for movie scene recognition and how to process a movie clip before classifying it.

Data Preparation

As this research report project progresses, custom datasets are built for testing and prototyping. I begin with a straightforward, limited custom dataset of YouTube movie clips and data further divided into four categories of Action, Horror, Romantic and Sci-Fi. The custom dataset is created from different 20 action scenes, 20 horror scenes, 20 Sci-Fi scenes and 20 romantic scenes. The custom dataset has around 2693 images for testing and 10771 images for both training categories. With this dataset, the network will be tested quickly in order to determine its correctness and make quick adjustments to its parameters.

Datasets

- Number of Categories: 4
- Total Scene Videos:
    - 20 action videos
    - 20 horror videos
    - 20 sci-fi videos
    - 20 romantic videos.
- Source of Videos: YouTube
- File Format: Videos (.Mp4) & Images (.JPG)
- Total Number of Frames in a Single Category: 3,600

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about



Figure 2 Movies different Scenes Dataset Example Frames

Data Pre-Processing

Simple data pre-processing is used in this research; 10% of the movies frames as image is set aside as a validation set and another 20% is set as a testing set for the dataset used to train the network. The validation set is used in this instance together with the training dataset to build the network and track the training results. It acts as an indication of network convergence. The testing set remains available for use after training as a measurement of recognition accuracy. Additionally, additional movie sequences and trailers will be utilized to test the network's generalization capabilities.

All of the frames are initially collected from videos from a single category at 30 frames per second (FPS) and labelled, after which training and testing sets are randomly divided. The frames are scaled down in this procedure to 224 × 224 resolution for standardization and simplicity.

Proposed Method

This section discusses the flow diagram of proposed method. The flow diagram of the proposed method is shown in fig 3. Data preparation and data pre-processing are discussed in the previous section respectively.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

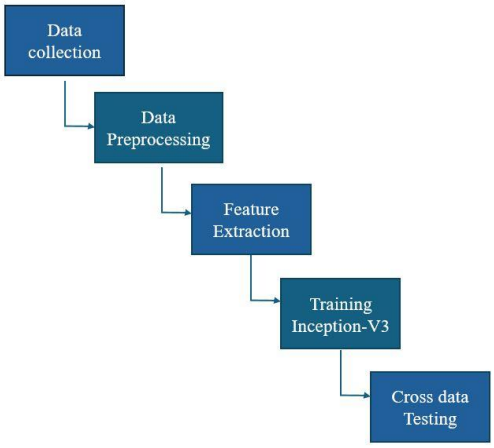


FIGURE 3 Flow Diagram of Proposed Method

Proposed Keyframes Selection Model

This research build upon the approach by Hussain et al [19] for efficiently and accurately removing redundant video frames by leveraging popular supervised learning-based image feature extraction methods in the field of computer vision. The similarity measure serves as an indicator of the degree to which two frames are similar to each other. By measuring the distance between dimensions that describe the characteristics of an object, the closeness of the object can be determined. [16] In this context, frames or images with small differences in similarities indicate a higher level of similarity, while those with large differences indicate a lower level of similarity. This standard distance measurement method plays a crucial role in determining which video frames should be retained or discarded.

The recognition of Content-Based Frame Retrieval (CBFR) in videos has been extensively studied and achieved through various feature detectors and descriptors, which are compared against each other in the existing literature [15]. By employing Content-Based Image Retrieval (CBIR) techniques, it becomes possible to identify and eliminate redundant video frames from captured video streams. In some circles, CBIR is also referred to as image-based search [16]. The term CBIR is used when a combination of feature descriptors, similarity measure techniques, and computer vision algorithms is employed to enhance image recognition. The proposed approach involves converting videos from different classes into frames, followed by the identification of redundant frames using a

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

threshold value applied to each individual frame. CNN features are then learned and extracted from the video frames using the Inception-v3 model, which provides a more accurate representation. As part of the feature extraction process using the Inception-v3 model, the images are resized to meet the model's input layer requirements. Specifically, the standard Inception-v3 CNN model necessitates 299x299 RGB images as input [17].

Several state-of-the-art dissimilarity methods have gained popularity in this domain, including the Euclidean Distance, the Hamming Distance, and the Cosine Similarity. The similarity score, often referred to as the degree of similarity, is an unsupervised learning method that quantifies the level of similarity between two data points. A similarity score of zero indicates virtual identity, signifying that the data points are completely identical.

Similarity Measure Using Euclidean Distance

When two points in Euclidean space are separated by an ordinary straight line, the distance between them is known as the Euclidean Distance or the Euclidean Metric [9]. It is defined as follows: two data point will be considered the same, if the distance between the two data points in an Euclidean space is zero, greater distance represents big difference between the data points (frames). The Pythagorean metric is a term that has been used to refer to the metric in older literature to describe it. The Euclidean norm, also known as the L2 norm or the L2 distance, is a shortened term for the Euclidean distance. For getting the basic understanding of distance metric (L2 distance) we can go through an example of distance measurement between two frames represented by two features vectors (features-1 and features-2).

Similarity = if the distance of X, Y <=15 (Where X, Y are two distinct video frames)

Features_1 = [34, 38, 41]

Features_2 = [36, 31, 25]

$$\text{Distance} = \sqrt{(x_i - y_i)^2 + (x_{i+1} + y_{i+1})^2 + \cdots + (x_n + y_n)^2}$$

In the fig 4, the key frames selection methodology is explained. To check a video frame whether it is keyframe or redundant frame, first of all the CNN features from two sequential frames are extracted. The Euclidean distance is a similarity measure technique used to check the similarity or difference between two data points. The

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

two features' vectors are passed to the Euclidean distance method to calculate the degree of similarity. Based on the Euclidean distance, the decision is taken whether to keep the frame as keyframe or discard it.

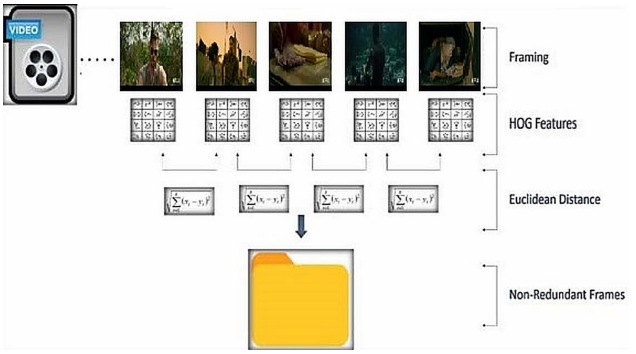


Figure 4 RVFR using CNN and Similarity measure [19]

In machine learning, the similarity measure, or Euclidean metric, is an innovative method for calculating the distance between data points in a 2D space. This technique is frequently utilized in computer vision and text processing to determine the degree of similarity between two images or documents. Similarity indicates the degree to which two frames are comparable. The similarity metric is the distance between the dimensions that represent the properties of objects. If the similarity distance is small, it indicates that the frames/images are highly similar, while a larger distance indicates that they are less similar. The popular method for calculating distance is the Euclidean distance which is used to calculate the degree of similarity between frames of a video in order to select or eliminate a frame. Using the Euclidean distance, also known as the Euclidean metric, to compare two objects or data points the Euclidean method figures out how far apart two things are. The smaller the distance, the more similar the things are.

The proposed key frames section technique will permit the frames containing unique content to be made accessible, while frames containing redundant content in the dataset will be eliminated. Data from the pretrain network will replace the data of the key frames. Before training the model on the custom dataset, the old labels, features, and weights must be removed. The hypermeter and the keyframes dataset will be added to the CNN model so that it can be retrained on the two-class violence dataset.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Inception-V3 Transfer Learning Model

Transfer learning is a machine learning technique that involves an existing pre trained model for one task as the foundation used to create a new model for another task. Consequently, the data utilized by the first model is transferred to the second model and the term "transfer learning" has its roots in human psychology [5].

Convolutional Neural Networks (CNNs) are specialized models that have been developed specifically for processing of image and video data. Alex net was the first model to be selected as the winner of the ImageNet competition. ImageNet is a benchmark dataset that contains 1000 classes and is used to validate a CNN model, which is a convolutional neural network. Because convolutional neural networks are ranked as deep learning models, a large amount of data must be collected and analyzed in order to train a deep learning classification model with convolutional neural networks. CNN, RNN, and other deep learning models perform poorly on small datasets. The CNN models are divided into two portions, with each portion for a specific task within the overall framework. The input layer, some hidden layers, and a final classification output layer are the same as in other neural network models. CNNs are no different than other neural network models in this regard.

The CNN models have been pre-trained, which means that they have gone through the process of learning from a dataset of 1000 class images, known as ImageNet, before being made available. CNN training is actually the process of calculating the optimal weights for the features extracted from the images, which is a mathematical procedure. For this study, the last three layers of this MATLAB's CNN model are removed and as a result of this removal, all of the old data associated with this part of network is also removed. This includes features, weights, labels, and bias values. Therefore, a newest of three layers have been added to the inception-v3 models: a fully connected layer, a SoftMax layer, and a classification layer. Various other pieces of information, such as parameters and hyper-parameters, are adjusted in accordance with the findings of the analysis. An optimization algorithm is used to determine the optimal learning rate for the CNN model in order to maximize its performance. It is possible to achieve optimal learning by calculating the most optimal weights for the image features and selecting the most appropriate learning strategy. Some standard layers that are used in almost all CNN models and are listed

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

in no particular order are Convolutional, Pooling, Normalization, ReLU, Sigmoid, Fully Connected and SoftMax for classification.

Proposed Methodology Using Inception V3 Model

The proposed model is deep learning-based system that will movie categories using supervised learning approach. The first phase of any video preprocessing system is framing, in which video data is converted to video frames where each single frame is a 2D RGB Image

Custom movies scenes datasets are used in this research for proposed model development and performance evaluation. For this purpose, we have employed a feature descriptor. The inception-v3 transfer learning model is implemented in order to extract robust features from every video frame. This means that we started with a model trained on a different problem. A CNN model consists of two components: the features extractor and classification. CNN has multiple layers, with distinct operations performed at each layer to extract high-dimensional image features. CNN is a global feature descriptor, which means that it processes the entire image to extract features, as opposed to first identifying points of interest and then extracting features.

Cross-validation is a mechanism used in statistics and machine learning that involves resampling a dataset in order to generate two subsets, one for training, and one for validation, each with a specific percentage of the original data. In this particular research simulation, the hold-out cross validation method, which is also known as the percentage-based splitting approach, is utilized with a standard split percentage value of 80:20 percent, which is chosen at random.

CNNs are ideal for identifying and categorizing characteristics of image and video data. In our application for detecting violence in video data, each frame can be represented by CNN's features, and sequential information can be extracted relatively easily using a CNN model. The scenario of feature representation after the original images are resized to a resolution of 224x224x3. The CNN classification layers are used for classification, and the Inception-v3 CNN model is used for feature extraction. The two components of the model are classification and feature extraction. Transfer learning only employs the CNN model's features extraction layer, leaving out the irrelevant fully connected, SoftMax, and classification output

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

layers. The table presents the model's inception-v3 structure.

The trained transfer learning model is a set of validated data on movies scenes. The validation set is created using the cross-validation method, and 20 percent of the test set's data is available. Each frame is transmitted to the trained network, classifying the image and assigning a label to each frame based on the category. After the validation phase is complete, the predicted class labels are compared with the actual class labels to determine which video frames were correctly and incorrectly classified. The final output of the classification will be either action, horror romantic and Sci-Fi video frame. The maximum predicted of certain category will be our recommended movie category. The validation set consists of image data and their corresponding labels. The confusion matrix is a frequently employed technique for distributing the results of correct and incorrect classifications. In addition, it contains information regarding the TP, FP, FN, and TN metrics used to evaluate the model's performance. Four metrics are used to evaluate the model's performance: accuracy, precision, recall, and f-measure.

RESULTS

Classification on Custom Movies Scenes Dataset

Given the simplicity of the categorization and the straightforward division of the dataset into three sections training (70%), validation (10%), and testing (20%) the network is trained for 100 iterations. Table 1 provides information about the dataset.

Table 1:

Dataset	Custom Scenes
Training image per category	Over 2500
Validation image per category	Over 250
Testing image per category	Over 500
Cross dataset testing image	100

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Classification Report

Table 2

Category	Precision	Recall	F1-Score
Action 0	0.44	0.64	0.52
Horror 1	0.74	0.68	0.71
Romantic 2	0.83	0.66	0.74
Sci-Fi 3	0.57	0.57	0.57
Macro avg	0.64	0.64	0.63
Weighted avg	0.67	0.64	0.65

Analysis of Custom Movies Scenes Classification Result

The baseline model achieved 0.66 accuracy, while the fine-tuned model improved slightly to 0.7225. This modest gain is expected because both training and testing data are taken from frames of a recently released movie trailer, with the test set containing scenes not present in the training set. However, when applying the same network to video clips rather than isolated frames, the additional temporal context may help improve accuracy further.

After Model Predict the scene category, the number of maximums predicted in single category which shown in figure 3 & 4.

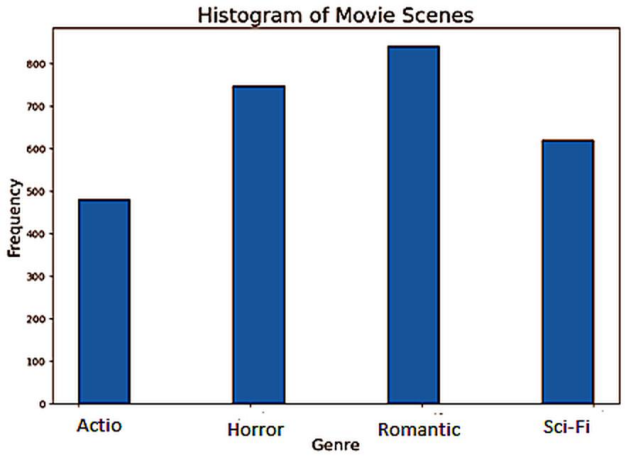


Figure 5 Bar Chart of Movie Scenes Classification

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

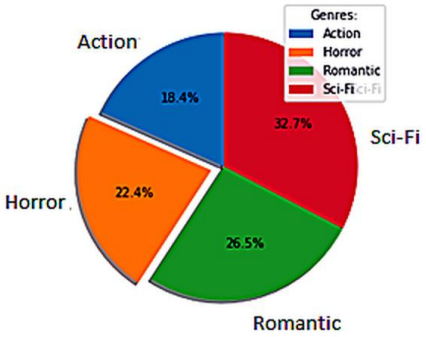


Figure 6 Pie Chart of Movie Scenes Classification

Confusion Matrix

Table 3

	Action 0	Horror 1	Romantic 2	Sci-Fi 3
Action 0	126	2	5	1
Horror 1	23	79	3	17
Romantic 2	48	7	79	9
Sci-Fi 3	36	30	5	83

Receiver Operating Characteristics (Roc) With Multi Class

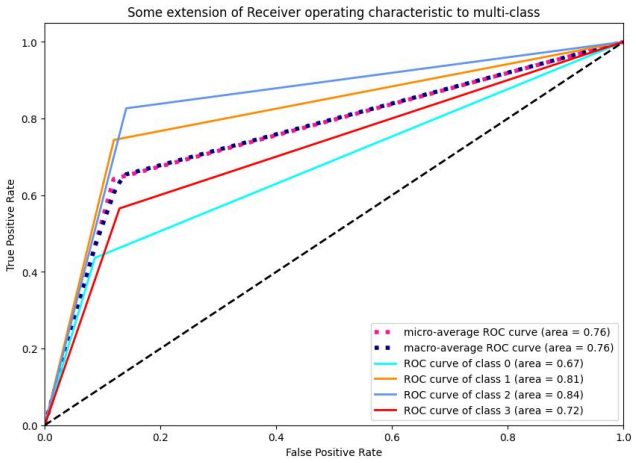


Figure 7 ROC with Multi Class

The findings of this research highlight the potential of leveraging deep learning techniques, specifically the Inception-V3 Convolutional Neural Network (CNN) model combined with transfer learning, to address the challenges of movie scene

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

recognition and genre classification. The proposed methodology demonstrates promising results in classifying movie scenes into four distinct genres—action, horror, romantic, and sci-fi—with an overall accuracy of 72%, as indicated in the classification report. This discussion elaborates on the implications of the results, the strengths and limitations of the proposed approach, and potential avenues for future research.

Interpretation of Results

The classification report (Table 2) reveals that the model performs reasonably well across all four genres, with precision, recall, and F1-scores ranging from 0.44 to 0.83. The romantic genre achieved the highest precision (0.83) and F1-score (0.74), suggesting that the model is particularly effective at identifying romantic scenes with minimal false positives. In contrast, the action genre exhibited the lowest precision (0.44) and F1-score (0.52), indicating potential challenges in accurately distinguishing action scenes from other genres. The macro-average F1-score of 0.63 and weighted-average F1-score of 0.65 further suggest that the model maintains a balanced performance across the genres, despite variations in individual class performance.

The accuracy improvement from 0.66 to 0.7225 when transitioning from a simple model to a fine-tuned model underscores the efficacy of transfer learning in enhancing classification performance on a custom movie scene dataset. This improvement is particularly notable given that the testing dataset comprised unseen movie trailer frames, which likely introduced additional complexity due to variations in scale, rotation, illumination, and translation challenges highlighted in the problem formulation. The use of Inception-V3 for feature extraction likely contributed to the model's ability to capture high-level visual representations, enabling robust genre classification despite these non-invariant features.

Strengths of the Proposed Approach

The proposed methodology offers several key strengths. First, the use of transfer learning with the Inception-V3 model allowed for efficient training on a relatively small custom dataset by leveraging pre-trained weights from the ImageNet dataset. This approach significantly reduced training time while maintaining high accuracy, addressing the computational challenges associated with large-scale movie scene

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

datasets. Second, the keyframe selection technique, based on Euclidean distance as a similarity measure, effectively reduced redundant frames, thereby optimizing the dataset for training and improving model efficiency. Third, the development of a GUI-based application enhances the practical applicability of the model, providing a user-friendly interface for video analysis and visualization of scene distributions. This feature has potential implications for real-world applications, such as content-based video recommendation systems and targeted advertising.

The custom movie scene dataset, comprising over 2500 training images per category, represents a significant contribution to the field. By constructing a dataset tailored to the specific task of genre classification, this research addresses the limitations of generic datasets that may not fully capture the nuances of movie scenes. The use of cross-validation with an 80:20 split further ensured robust evaluation, minimizing overfitting and providing reliable performance metrics.

Limitations and Challenges

Despite its strengths, the proposed approach has several limitations. The moderate recall for the action genre (0.64) suggests that the model struggles to identify all relevant action scenes, potentially due to the high variability in action sequences, which may include rapid scene changes, diverse camera angles, or blurred backgrounds. This aligns with the challenges noted in the literature review, where movie shots often prioritize close-ups of characters, making background-based scene recognition difficult. Additionally, the dataset, while substantial, may not fully capture the diversity of movie scenes across different directors, styles, or production qualities, which could limit the model's generalizability.

The reliance on Euclidean distance for keyframe selection, while effective, may not fully account for complex semantic differences between frames. Alternative similarity measures, such as Cosine Similarity or Hamming Distance, could potentially improve keyframe selection by capturing more nuanced differences in visual content. Furthermore, the model's performance was evaluated on a single-label classification task, whereas real-world movie scenes often exhibit multi-label characteristics (e.g., a scene may be both action-packed and romantic). Extending the model to handle multi-label classification could enhance its applicability to more complex scenarios.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

IMPLICATIONS AND APPLICATIONS

The results of this research have significant implications for both academic research and industry applications. By addressing the limitations of traditional machine learning algorithms in extracting non-invariant features, this study contributes to the advancement of computer vision techniques for video analysis. The proposed model's ability to classify movie scenes with reasonable accuracy paves the way for more personalized content recommendation systems, which can complement collaborative filtering approaches by incorporating content-based features. The GUI-based application further enhances the practical utility of the model, enabling stakeholders such as streaming platforms, advertisers, or content creators to analyze and categorize video content efficiently.

Beyond entertainment, the techniques developed in this research have potential applications in other domains, such as surveillance video analysis (e.g., identifying specific events in CCTV footage) or traffic monitoring (e.g., detecting anomalies in traffic camera feeds). The semantic understanding of video content achieved through CNN-based scene recognition could also inform the development of more sophisticated video search engines or content-based image retrieval systems.

CONCLUSION AND FUTURE WORK

This research article attempts to address the question of whether and how scene information may be extracted from a film.

This research demonstrates the effectiveness of using the Inception-V3 CNN model with transfer learning for movie scene recognition and genre classification. The proposed methodology achieves a promising accuracy of 72% on a custom movie scene dataset, highlighting the potential of deep learning techniques to overcome the challenges of non-invariant feature extraction. While limitations remain, particularly in the classification of action scenes, the proposed approach offers significant contributions to the field of computer vision and video analysis. The development of a GUI-based application further enhances the practical applicability of the model, paving the way for future advancements in content-based video recommendation and analysis systems. By addressing the identified limitations and exploring the suggested research directions, future studies can build on these findings to develop more robust and versatile video analysis tools.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

A particular movie scene dataset, created by the author [6], and movie frames with scene labels make up the MovieScene dataset, which has four categories with at least 2600 training pictures for each category. This research uses the currently popular Convolutional Neural Network technique for scene detection.

In addition to serving as the benchmark for classifying video clips, the single image classification test result illustrates the network's capacity to identify a scene inside an image and facilitates comparison with other networks. Due in part to variations in the training photos within each category, such as the density and diversity of the data, the classification results vary greatly between categories. However, by fine-tuning the model, the network's accuracy rises by over 10%.

A significant amount of misclassification occurs when the input image is not ideal for classification, such as a fuzzy backdrop or an actor taking up too much area in a single frame, according to tests done on movie clips. In an attempt to exclude some of those frames, a category of likely faces is added to the dataset.

It should be mentioned that the accuracy presented in this report may not accurately represent the overall performance on a large movie dataset because video testing is only done on a small number of films and segments due to the high cost of human resources. Additionally, human error is a problem here because my project mates and I conduct the testing at the institution.

The approach and procedures used in this project may be extended to future video feature analysis, making this research report a successful step in the direction of movie content extraction. Additionally, the project's output may be used to apply basic tags to films with a dominant scene type. For example, the film "Extraction" has the greatest percentage of "Action" tags and the second-highest percentage of "Sci-Fi" tags. In the future, the classifier-generated tags may potentially be used as search terms to enhance the movie search experience.

The answer to the question is that, with a certain degree of accuracy, movie scene information may be extracted by classification using a CNN. The techniques shown in this research are applicable to general movie analysis, such as identifying a film's dominant scene. To improve its ability to forecast people's tastes, this data may also be fed into a movie recommendation engine. But this network might not be able to handle more in-depth frame-by-frame analysis of a film.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

FUTURE WORK

Numerous studies, such as [7], have demonstrated that a CNN's effectiveness is largely dependent on the training dataset; a high-quality, representative dataset may be very beneficial. In order to create a scene dataset based solely on movie frames rather than real-world situations, it would be beneficial to gather a large number of annotated movie scene photos.

Second, to improve performance, a far more intricate and deep network may be taught. Training and fine-tuning a deep neural network can take weeks, and many deep neural networks must be trained to obtain a parameter set that works effectively. This is one of the reasons a deeper neural network isn't trained for this study report; a well-known network structure like AlexNet [8] will take more than six months to install, not to mention the time needed for research.

Furthermore, a dataset's density and variety may be used to better gauge its quality; these quantitative metrics for assessing dataset biases and contrasting other datasets can be very beneficial for future studies.

While the focus of this research report project is on a scene-centric approach to the problem, an object-centric approach to movie scene identification is equally feasible and worthwhile to investigate. One potential benefit is that networks built for ImageNet, a considerably bigger object dataset, are already in existence.

REFERENCES

[1] Imdb. http://www.imdb.com/. Accessed: 2015-06-25.

[2] F. Sadikoglu and S. Uzelaltinbulat, "Biometric Retina Identification Based on Neural Network," Procedia Comput. Sci., vol. 102, no. August, pp. 26–33, 2016, doi: 10.1016/j.procs.2016.09.365.

[3] J. Liu, W. H. Choi, and J. Liu, "Personalized Movie Recommendation Method Based on Deep Learning," Math. Probl. Eng., vol. 2021, 2021, doi: 10.1155/2021/6694237.

[4] S. H. Min and I. Han, "Recommender systems using support vector machines," Lect. Notes Comput. Sci., vol. 3579, no. December 2005, pp. 387–393, 2005, doi: 10.1007/11531371_50.

[5] C. Szegedy, V. Vanhoucke, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2014.

Online ISSN 3007-3197
Print ISSN 3007-3189

[6] https://drive.google.com/drive/folders/1zoDzHhXA8xCPzL6aAeX52IiBvk55RM75?usp=share_link

[7] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Advances in Neural Information Processing Systems, pages 487–495, 2014

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[9] I. M. Hameed, S. H. Abdulhussain, B. M. Mahmmod, I. M. Hameed, S. H. Abdulhussain, and B. M. Mahmmod, "Content-based image retrieval : A review of recent trends COMPUTER SCIENCE | REVIEW ARTICLE Content-based image retrieval : A review

[10] Singh, Vineeta, Deeptha Girish, and Anca L. Ralescu. "Image Understanding-a Brief Review of Scene Classification and Recognition." *MAICS* 2017 (2017): 85-91.

[11] D. Bose et al., "MovieCLIP: Visual Scene Recognition in Movies," 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2023, pp. 2082-2091, doi: 10.1109/WACV56688.2023.00212

[12] Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadrana, "Content-Based Video Recommendation System based on stylistic visual features," *Journal on Data Semantics*, vol. 5, no. 2, pp. 99–113, Feb. 2016, doi: 10.1007/s13740-016-0060-9.

[13] P. Khadka and P. Lamichhane, "Content-based recommendation engine for video streaming platform," *arXiv.org*, Aug. 16, 2023. https://arxiv.org/abs/2308.08406

[14] Pazzani, M.J., Billsus, D. (2007). Content-Based Recommendation Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds) The Adaptive Web. Lecture Notes in Computer Science, vol 4321. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_10

Online ISSN 3007-3197
Print ISSN 3007-3189

[15] Reddy, S., Nalluri, S., Kunisetti, S., Ashok, S., Venkatesh, B. (2019). Content-Based Movie Recommendation System Using Genre Correlation. In: Satapathy, S., Bhateja, V., Das, S. (eds) Smart Intelligent Computing and Applications . Smart Innovation, Systems and Technologies, vol 105. Springer, Singapore. https://doi.org/10.1007/978-981-13-1927-3_42

[16] N. J. B. Dias, G. T. Laureano, and R. M. Da Costa, "Keyframe selection for visual localization and mapping Tasks: A Systematic Literature review," *Robotics*, vol. 12, no. 3, p. 88, Jun. 2023, doi: 10.3390/robotics12030088.

[17] J. Gao, J. Yang, J. Zhang and M. Li, "Natural scene recognition based on Convolutional Neural Networks and Deep Boltzmannn Machines," 2015 IEEE International Conference on Mechatronics and Automation (ICMA), Beijing, China, 2015, pp. 2369-2374, doi: 10.1109/ICMA.2015.7237857.

[18] S. El-Bana, A. Al-Kabbany, and M. Sharkas, "A multi-task pipeline with specialized streams for classification and segmentation of infection manifestations in COVID-19 scans," PeerJ Computer Science, vol. 6, p. e303, Oct. 2020, doi: 10.7717/peerj-cs.303

[19] A. Hussain, "An efficient video frames retrieval system using speeded up robust features based bag of visual words," ADCAIJ ADVANCES IN DISTRIBUTED COMPUTING AND ARTIFICIAL INTELLIGENCE JOURNAL, vol. 12, p. e28824, Dec. 2023, doi: 10.14201/adcaij.28824.