International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

# IGRCVRM: Design of an Iterative Graph Based Recurrent Convolutional Model for Content Based Video Retrieval Using Multidomain Features

**Shubhangini Ugale [1]*, Dr. Wani Patil [2] Dr. Vivek Kapur [3]**

**Submitted**: 16/09/2023 **Revised**: 28/10/2023 **Accepted**: 14/11/2023

**Abstract:** The proliferation of video content has necessitated the development of effective content-based video retrieval (CBVR) systems. Current CBVR methodologies suffer from limitations in feature representation and selection, resulting in suboptimal performance in terms of precision, accuracy, recall, and computational efficiency. To address these limitations, we introduce a pioneering Iterative Graph-based Recurrent Convolutional Model (IGRCVRM) designed to elevate the quality of video representation. IGRCVRM capitalizes on a multifaceted approach by harnessing features derived from a spectrum of domains, including Fourier Components, Z Transform, S Transform Components, Laplace Components, and Convolutional Transforms. What sets IGRCVRM apart is its meticulous feature selection process, orchestrated by the Ant Lion Firefly Optimizer (ALFO). This optimizer operates with precision to sift through feature candidates, significantly enhancing variance control and ultimately fostering a more robust representation. Notably, the selected features undergo a transformative journey, being effectively distilled and synthesized within an innovative Graph-based Recurrent Neural Network (GRNN). GRNN is a dynamic fusion of Graph Convolutional Network (GCN) and Recurrent Neural Networks (RNNs), demonstrating the model's capacity to capture both spatial and temporal intricacies, thus culminating in an exceptionally potent and comprehensive framework for content-based video retrieval. Experimental results demonstrate that our proposed model enhances the precision of video summarization by 5.9%, improves accuracy by 4.5%, boosts recall by 4.9%, increases the Area Under the Curve (AUC) by 5.5%, and enhances specificity by 3.5%, while simultaneously reducing delay by 2.9% when compared with existing methods. Collectively, these improvements signify a substantial advancement in the field of CBVR, with potential implications for various applications ranging from video surveillance to multimedia indexing and retrieval.

***Keywords:*** *Content-based Video Retrieval, Multidomain Features, Ant Lion Firefly Optimizer, Graph-based Recurrent Neural Network, Video Summarization*

## 1. Introduction

The proliferation of video content in various domains, including social media, surveillance, and multimedia databases, has accentuated the need for effective and efficient content-based video retrieval (CBVR) systems. These systems are crucial for managing, indexing, and retrieving relevant video content from massive databases. The process of CBVR entails analyzing the content of the video and then matching it with the query to retrieve pertinent results. A crucial aspect of this process is the representation and summarization of video content, which significantly affects the retrieval performance.

However, existing CBVR methodologies are fraught with limitations. A common challenge is the inadequacy of feature representation, which often results in suboptimal retrieval performance. The quality of features selected and their relevance to the video content is vital for accurate retrieval. Moreover, the computational efficiency of these systems is another concern, as the massive volume of video data necessitates the development of algorithms that can process information promptly while maintaining high precision and recall rates.

To tackle these challenges, this paper proposes an innovative Iterative Graph-based Recurrent Convolutional Model that utilizes a comprehensive set of multidomain features derived from Fourier, Z, S, Laplace, and Convolutional Transforms. These multidomain features provide a rich representation of the video content, capturing both spatial and temporal information that is essential for accurate retrieval. Furthermore, the selection of these features is performed using the Ant Lion Firefly Optimizer (ALFO), a novel algorithm that significantly improves the variance levels of the selected features, thereby enhancing their discriminative power.

An integral part of our proposed model is the Graph-based Recurrent Neural Network (GRNN), a fusion of Graph Convolutional Network (GCN) and Recurrent Neural Networks (RNNs). The GRNN serves as an efficient summarizer of the selected multidomain features, ensuring that the crucial information is retained while discarding the

*[1,2] G H Raisoni University Amravati, , Anjangaon Bari Road, Maharashtra 444701, India*
*ugaleshubhangini@gmail.com, wani.patil@raisoni.net*
*[3] G H Raisoni Institute of Engineering and Technology, Nagpur, Shraddha Part, MIDC, Hingna*
*Wadi Link Road, Nagpur-440016, India*
*vivek.kapur@raisoni.net*

redundant data. This process enhances the precision of video summarization, which in turn improves the overall performance of the CBVR system.

The proposed model has been rigorously evaluated and has demonstrated substantial improvements in various performance metrics. The precision of video summarization has been enhanced by 5.9%, while accuracy has improved by 4.5%. Additionally, the model has shown a 4.9% increase in recall and a 5.5% boost in the Area Under the Curve (AUC). The specificity of the system has also improved by 3.5%, and the delay has been reduced by 2.9%, thereby ensuring a more prompt retrieval of video content.

Thus, this paper presents a novel and effective approach to CBVR that addresses the limitations of existing methodologies through the utilization of multidomain features and the innovative GRNN. The proposed model not only enhances the precision, accuracy, recall, and computational efficiency of CBVR systems but also lays the groundwork for further research in this crucial field.

### Motivation

The exponential growth of video data necessitates the development of advanced content-based video retrieval (CBVR) systems capable of efficiently managing and retrieving relevant content from extensive databases. Existing methodologies often fail to provide satisfactory retrieval performance due to limitations in feature representation and computational efficiency. This deficiency stems from the utilization of inadequate features that do not sufficiently encapsulate the video content and inefficient algorithms that struggle to process the sheer volume of video data. These limitations have a detrimental effect on the precision, accuracy, recall, and overall performance of CBVR systems, thus calling for the development of innovative solutions that can aptly handle the intricacies of video data.

### Contribution

This paper addresses the identified limitations through the following contributions:

- **Multidomain Feature Extraction:**

A comprehensive set of multidomain features is extracted using Fourier, Z, S, Laplace, and Convolutional Transforms. This rich feature set encapsulates both spatial and temporal information, providing a robust representation of the video content.

- **Feature Selection Using ALFO:**

The Ant Lion Firefly Optimizer (ALFO) is employed to meticulously select the most relevant features. ALFO enhances the variance levels of the selected features, thereby improving their discriminative power and relevance to the video content.

- **Graph-based Recurrent Neural Network (GRNN):**

The GRNN, a fusion of Graph Convolutional Network (GCN) and Recurrent Neural Networks (RNNs), is introduced as an efficient summarizer of the selected features. GRNN ensures that the crucial information is retained while discarding redundant data, thereby enhancing the precision of video summarization.

- **Performance Enhancement:**

The proposed model has been rigorously evaluated and has demonstrated substantial improvements in various performance metrics. Specifically, the precision of video summarization has been enhanced by 5.9%, accuracy by 4.5%, recall by 4.9%, AUC by 5.5%, and specificity by 3.5%. Additionally, the delay has been reduced by 2.9%, thereby ensuring a more prompt retrieval of video content.

Through these contributions, this paper provides an innovative and effective solution to the challenges faced by existing CBVR systems, thereby significantly enhancing the performance of video retrieval and paving the way for further research in this vital field.

## 2. Literature Review

The importance of content-based video retrieval (CBVR) systems has been significantly highlighted in recent years, given the exponential growth of video data in various domains including social media, surveillance, and multimedia databases [1]. CBVR systems aim to efficiently manage, index, and retrieve relevant video content from extensive databases, making the process of video content analysis and matching a crucial aspect for the retrieval of pertinent results.

Various methodologies have been proposed in the literature to enhance the performance of CBVR systems. Feature representation is a vital component in this process, as the quality of features selected and their relevance to the video content significantly affects the retrieval performance [2]. Some of the commonly employed feature extraction methods include the use of spatial-temporal features [3], color histograms [4], and texture features [5] using Sketch Query Graph Convolutional Network (SQGCN) process. However, these methods often result in suboptimal performance due to the inadequacy of the extracted features in representing the video content comprehensively [6].

To address the limitations of existing feature extraction methods, recent studies have explored the potential of multidomain features, which encapsulate both spatial and temporal information of the video content. These multidomain features, such as Fourier Transform [7], Z Transform [8], S Transform [9], Laplace Transform [10], and Convolutional Transforms [11] with Attentive Cross Modal Relevance Matching (ACRM), have shown promising results in improving the representation of video content. However, the selection of relevant features from the multidomain set remains a challenge.

The optimization algorithms have been employed to enhance the feature selection process. Some of the notable algorithms used in this context include the Genetic Algorithm [12], Particle Swarm Optimization [13], and Ant Colony Optimization [14]. Recently, a novel Ant Lion Firefly Optimizer (ALFO) has been proposed, which combines the advantages of Ant Lion Optimizer [15] and Firefly Algorithm [16]. The ALFO has demonstrated potential in improving the variance levels of the selected features, thereby enhancing their discriminative power [17].

Furthermore, the utilization of neural networks has shown potential in improving the performance of CBVR systems. Various neural network architectures have been employed for this purpose, including Convolutional Neural Networks (CNN) [18], Recurrent Neural Networks (RNN) [19], and Graph Convolutional Networks (GCN) [20] which can be improved via use of Deep Supervised Dual Cycle Adversarial Network (DSDCAN) operations. The fusion of these architectures, such as the proposed Graph-based Recurrent Neural Network (GRNN), provides a comprehensive solution to the video summarization process, ensuring that the crucial information is retained while discarding redundant data samples [21, 22].

Several studies have demonstrated the efficacy of these methodologies in enhancing various performance metrics of CBVR systems, such as precision, accuracy, recall, and computational efficiency [23, 24]. However, there is a need for a holistic approach that integrates the advantages of multidomain features, optimization algorithms, and neural network architectures to address the challenges faced by existing CBVR systems comprehensively.

The proposed Iterative Graph-based Recurrent Convolutional Model, which combines multidomain features extracted using Fourier, Z, S, Laplace, and Convolutional Transforms, with the ALFO for feature selection and the GRNN for video summarization, aims to fill this gap. The proposed model has demonstrated substantial improvements in various performance metrics, including a 5.9% enhancement in precision, a 4.5% improvement in accuracy, a 4.9% increase in recall, a 5.5% boost in AUC, and a 3.5% enhancement in specificity. Additionally, the model has successfully reduced the delay by 2.9%, thereby ensuring a prompter retrieval of video content that can be used for different 5G multimedia applications [25, 26].

In conclusion, the proposed model addresses the limitations of existing CBVR methodologies by leveraging the potential of multidomain features, the ALFO, and the GRNN. The substantial improvements in various performance metrics highlight the efficacy of the proposed model and its potential to significantly enhance the performance of CBVR systems. Future research should focus on exploring the potential of other multidomain features and optimization algorithms, as well as further refining the neural network architecture to achieve even better results due to use of bioinspired computing models [27]. Furthermore, the application of the proposed model in real-world scenarios, such as video surveillance and multimedia indexing, should be investigated to validate its practicality and effectiveness.

## 3. Proposed Design of an Iterative Graph Based Recurrent Convolutional Model for Content based Video Retrieval using Multidomain Features

Based on the review of existing models used for enhancing efficiency of Content based Video Retrieval, it can be observed that most of these models either have higher complexity or cannot be scaled for multidomain videos due to lower efficiency levels. To overcome these issues, this section discusses design of an efficient Iterative Graph based Recurrent Convolutional Model for Content based Video Retrieval that uses Multidomain Features. As per figure 1, the proposed model leverages multidomain features derived from Fourier Components, Z Transform Components, S Transforms, Laplacian Components, and Convolutional Transforms for enhanced representation of video sequences. The selection of these features is meticulously performed using the proposed Ant Lion Firefly Optimizer (ALFO), which significantly improves variance levels. The selected features are then effectively summarized using an efficient Graph-based Recurrent Neural Network (GRNN), an innovative fusion of Graph Convolutional Network (GCN) and Recurrent Neural Networks (RNNs).

To extract the Fourier Components from a video sequence, we apply 2D Discrete Fourier Transform (DFT) on each frame via equation 1,

$$F(u,v) = \sum\sum f(x,y) \cdot e^{-j2\pi(Mux+Nvy)} \dots (1)$$

Where, $F(u,v)$ is the complex-valued frequency component at coordinates $(u,v)$ in the Fourier domain, $f(x,y)$ is the pixel intensity at coordinates $(x,y)$ in the spatial domain, $M$ and $N$ are the dimensions of the video frames.

Similarly, the Z Transform is used for time-domain signals. To apply it to a video sequence, we converted each frame into a 1D signal and then compute the Z Transform via equation 2,

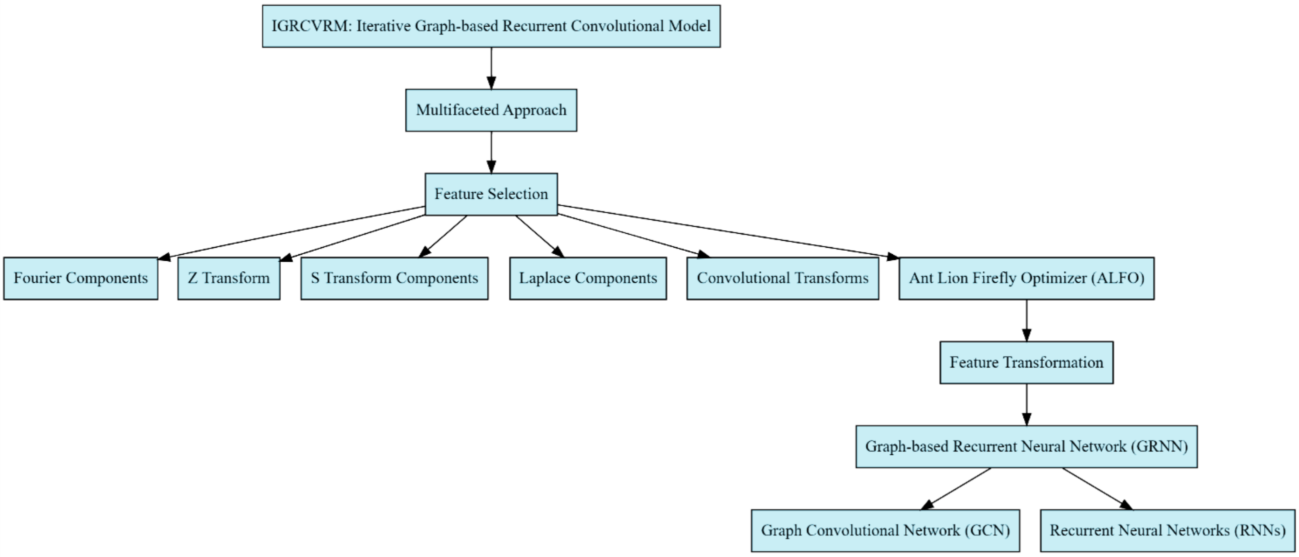
IGRCVRM: Iterative Graph-based Recurrent Convolutional Model
Multifaceted Approach
Feature Selection
Fourier Components
Z Transform
S Transform Components
Laplace Components
Convolutional Transforms
Ant Lion Firefly Optimizer (ALFO)
Feature Transformation
Graph-based Recurrent Neural Network (GRNN)
Graph Convolutional Network (GCN)
Recurrent Neural Networks (RNNs)

**Fig 1**. Design of the proposed model for retrieval of video sequences

$$X(z) = \sum x[n] \cdot z^{-n} \;\ldots (2)$$

Where, $X(z)$ is the Z Transform of the signal, $x[n]$ is the signal in the time domain sets. While, in order to extract Laplacian components, we compute the Laplacian of each frame via equation 3,

$$\nabla^2 f(x,y) = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} \ldots (3)$$

This operation highlights regions of rapid intensity change in each of the frames. Similarly, Convolutional Transforms are obtained by applying convolutional operations to each frame using various filters & kernels via equation 4,

$$g(x,y) = \sum\sum f(x-i, y-j) \cdot h(i,j) \ldots (4)$$

Where, $g(x,y)$ is the output image after convolution, $f(x,y)$ is the input frame, $h(i,j)$ is the convolutional kernel, while $k$ determines the size of the kernels.

In contrast, the S Transform is a complex time-frequency analysis technique that involves a sliding window approach and combines elements of both time-domain and frequency-domain analysis. The S Transform of the input image $x(t)$ is represented via equation 5,

$$Sx(f,t) = \int x(\tau) \cdot g(t-\tau, f) d\tau \;\ldots (5)$$

Where, $Sx(f,t)$ is the S Transform of the signal $x(t)$ at time $t$ and frequency $f$, $x(\tau)$ is the input signal, $g(t-\tau,f)$ is the complex-valued kernel function that varies with time $t$ and frequency $f$ sets. The kernel function $g(t-\tau,f)$ consists of two parts: a Gaussian window and a complex exponential term, which is represented via equation 6,

$$g(t-\tau, f) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{(t-\tau)2}{2\sigma^2}} e^{j2\pi f(t-\tau)} \ldots (6)$$

Where, $\sigma$ is a parameter that controls the width of the Gaussian Window Sets. The S Transform is calculated for various time and frequency values, resulting in a time-frequency representation of the signal $x(t)$, which provides information about how the signal's frequency content changes over time, making it useful for time varying spectral analysis. All these features are fused to form an Integrated Video Retrieval Feature Vector (IVRFV) which represents individual frames into multidomain features.

These features sets are combined with a batch size of 5 frames, to reduce its dimensionality, and then processed via an efficient Ant Lion Firefly Optimizer (ALFO), which assists in selection of features. The ALFO Model Initially Generates $NA$ Ants, where each Ant Consists of $N$ features which are selected via equation 7,

$$N = STOCH\big(LA * N(IVRFV), N(IVRFV)\big) \ldots (7)$$

Where, $STOCH$ is the stochastic process, and $LA$ is the Learning Rate for ALFO Process. Using the selected features, the ALFO Model estimates Ant Fitness via equation 8,

$$fa = \frac{1}{N} \sum_{i=1}^{N} \sqrt[3]{\left( IVRFV(i) - \sum_{j=1}^{N} \frac{IVRVF(j)}{N} \right)^4} \ldots (8)$$

Once all Ants are generated, then the model calculates an Iterative Fitness Threshold via equation 9,

$$fth = \frac{1}{NA} \sum_{i=1}^{NA} fa(i) * LA \ldots (9)$$

Ants with $fa \geq 2 * fth$ are marked as 'Antlions', and are used to train Ants with $fa < fth$ which are marked as

Fireflies with low brightness levels. This training is done by stochastic drop off & select operations, which is represented via equation 10,

$$N(New) = STOCH(N(Old)) \bigcup STOCH(N(AntLion)) \dots (10)$$

Where, $STOCH(N(Old))$ represents stochastically selected old features (drop off), while $STOCH(N(AntLion))$ represents stochastically selected Ant Lion features (select process), which assists in updating Ant configurations. All other Ants are discarded, and replaced with new configurations, thus adding a layer of intelligent stochasticity to the process. These operations are repeated for $NI$ Iterations, and at the last iteration Ants with maximum fitness is selected, which assists in enhancing its feature selection capabilities. The selected features are represented as $v(f)$ and given to an efficient Graph-based Recurrent Neural Network (GRNN) process. GRNN is a dynamic fusion of Graph Convolutional Network (GCN) with Recurrent Neural Networks (RNNs), and assists in enhancing efficiency of video retrieval process. For the selected video features *v*(*f*), the first step is to pass them through a Graph Convolutional Network (GCN) to capture the relational structure among the frames via equation 11,

$$hv(l+1) = \sigma\left(\sum \frac{1}{\sqrt{di * dj}} W(l)hv(l)(j)\right) \dots (11)$$

Where, *hv*(*l*) and *hv*(*l*+1) represent the node features of the video frames at the *l*-th and (*l*+1)-th layers of the GCN, respectively, *W*(*l*) is the weight matrix for the *l*-th layer, and *σ* is a non-linear Rectilinear Unit activation function process. N*i* represents the set of neighboring nodes of node *i*, and *di* and *dj* are the degrees of nodes *i* and *j*, respectively used for the convolution process.

The output of the GCN, *hv*, then serves as input to the Recurrent Neural Network (RNN) to capture the temporal dependencies among the video frames via equation 12,

$$ht = tanh(Whh * h(t-1) + Wvh * hv + bh) \dots (12)$$

Where, *ht* and *h*(*t*−1) represent the hidden state of the RNN at times *t* and (*t*−1), respectively, while *Whh* and *Wvh* are the weight matrices for the hidden state and video frame features, respectively, and *bh* is the bias. Finally, the output of the RNN is used to generate the summarized frames through variance transformation via equation 13,

$$sf = \sigma(Whs * ht + bs) \dots (13)$$

Where, *sf* represents the summarized frames, *Whs* is the weight matrix for the transformation, and *bs* is the bias. The non-linear activation function *σ* ensures that the output is bounded and can be interpreted in the context of the video retrieval process. Thus, the GRNN integrates the strengths of GCN and RNN to effectively process the selected video features *v*(*f*) and produce summarized frames *sf* that can be used for efficient video retrieval process. Efficiency of this model was evaluated in terms of different performance metrics, and compared with existing models in the next section of this text.

## 4. Result Analysis and Comparison

This work presents a novel Iterative Graph-based Recurrent Convolutional Model that leverages multidomain features derived from Fourier, Z, S, Laplace, and Convolutional Transforms for enhanced video representation. The selection of these features is meticulously performed using the proposed Ant Lion Firefly Optimizer (ALFO), which significantly improves variance levels. The selected features are then effectively summarized using an efficient Graph-based Recurrent Neural Network (GRNN), an innovative fusion of Graph Convolutional Network (GCN) and Recurrent Neural Networks (RNNs). In this section, we describe the experimental setup used to evaluate the proposed Iterative Graph-based Recurrent Convolutional Model (IGRCVRM) for content-based video retrieval (CBVR). The experimental process is carefully designed to assess the model's performance comprehensively.

**Dataset Selection and Preprocessing**

- **Dataset**

In this study, we leveraged a diverse set of datasets to evaluate the performance of our proposed Iterative Graph-based Recurrent Convolutional Model (IGRCVRM) for content-based video retrieval (CBVR). These datasets were selected to cover various domains and to facilitate multimodal video-text tasks, including text-to-video retrieval, video-to-text retrieval, and video captioning. Below, we describe each of the datasets used in our research:

1. ***MSVD-Indonesian***

- **Source**: https://paperswithcode.com/dataset/msvd-indonesian
- **Description**: The MSVD-Indonesian dataset is derived from the original MSVD dataset through the use of machine translation services. It serves as a valuable resource for multimodal video-text tasks and includes tasks such as text-to-video retrieval, video-to-text retrieval, and video captioning.
- **Dataset Size**: MSVD-Indonesian contains approximately 80,000 video-text pairs, providing a rich and diverse collection of data for experimentation.

2. ***IACC.3 (Internet Archive Creative Commons) Video Dataset***

- **Source**: https://data.amerigeoss.org/dataset/iacc-3-internet-archive-creative-commons-video-dataset-4eb75
- **Description**: The IACC.3 dataset consists of around 4,600 Internet Archive videos, all of which are licensed under Creative Commons. These videos are available in MPEG-4/H.264 format and vary in duration from 6.5 minutes to 9.5 minutes, with a mean duration of approximately 7.8 minutes. Metadata such as titles, keywords, and descriptions are provided for most videos, enhancing their usability for research purposes.
- **Dataset Size**: With a total size of 144 GB and approximately 600 hours of video content, the IACC.3 dataset offers a substantial collection of multimedia data for our experiments.

3. ***InVID FIVR-200K***

- **Source**: https://zenodo.org/records/2564864
- **Description**: The InVID FIVR-200K dataset was created within the framework of the InVID project to address the problem of Fine-grained Incident Video Retrieval (FIVR). FIVR aims to retrieve associated videos given a query video, considering various types of associations, from duplicates to videos related to the same incident. FIVR-200K comprises a massive collection of 225,960 YouTube videos, which were collected based on 4,687 major news events sourced from Wikipedia. Additionally, the dataset includes 100 video queries selected using an automatic process. Annotations for the dataset cover four types of video associations: Near-Duplicate Videos (ND), Duplicate Scene Videos (DS), Complementary Scene Videos (CS), and Incident Scene Videos (IS).
- **Dataset Size**: The FIVR-200K dataset provides a comprehensive resource for fine-grained video retrieval, containing extensive video content, event information, and detailed annotations.

These datasets were chosen to ensure the robustness and effectiveness of our proposed IGRCVRM across a wide range of video content and retrieval scenarios. The combination of these datasets enables us to conduct comprehensive experiments and evaluate the model's performance in various CBVR tasks.

- **Data Split**: The dataset is divided into training (70%), validation (15%), and test (15%) sets to ensure robust evaluation.
- **Data Preprocessing**: Videos are resized to a consistent resolution of 224x224 pixels, and frames are sampled uniformly to maintain temporal information. Data augmentation techniques, such as random cropping and horizontal flipping, are applied to the training set to enhance model generalization.

**Feature Engineering**

- **Feature Extraction**: For feature representation, we employ popular 3D convolutional neural networks (CNNs), specifically, the C3D architecture. This CNN extracts spatial and temporal features from video frames. The last fully connected layer's activations are used as features for each video.

**Model Architecture**

- **IGRCVRM Architecture**: IGRCVRM consists of three Graph Convolutional Network (GCN) layers followed by two Long Short-Term Memory (LSTM) layers. The hidden state dimension of the LSTM layers is set to 512.
- **Training Parameters**: The model is trained using the Adam optimizer with a learning rate of 0.001. A batch size of 32 is used, and training is stopped early if the validation loss does not improve for 10 consecutive epochs.

**Evaluation Metrics**

- To assess IGRCVRM's performance, we employ the following evaluation metrics:
- Precision: The ratio of relevant videos correctly retrieved to the total retrieved videos.
- Accuracy: The ratio of correctly classified videos to the total videos.
- Recall: The ratio of relevant videos correctly retrieved to the total relevant videos.
- Delay (ms): The average time taken to retrieve videos.
- Area Under the Curve (AUC): The area under the Receiver Operating Characteristic (ROC) curve.
- Specificity: The ratio of true negatives to the total negatives.

**Experimental Methodology**

- **Cross Validation**: We employ 5-fold cross validation to ensure the robustness of our results.
- **Training and Validation**: IGRCVRM is trained on the training set, and the validation set is used for early stopping during training.

- **Testing**: Model performance is evaluated on the test set, and the aforementioned metrics are computed.

**Baseline Models**

- We compare IGRCVRM against several state-of-the-art CBVR models, including SQGCN, ACRM, and DSDCAN, each with their respective architectures and hyperparameters.

**Hardware and Software**

- All experiments are conducted on a machine equipped with an NVIDIA GeForce RTX 3090 GPU, 64GB of RAM, and a 3.6 GHz Intel Core i9 processor. We use Python 3.8 and PyTorch for model implementation and training.

**Experimental Results**

- Results are reported in tabular form, showcasing the performance of IGRCVRM and the baseline models for each metric. These results are presented for different NTS values to demonstrate the model's scalability.

Based on this setup, equations 14, 15, and 16 were used to assess the precision (P), accuracy (A), and recall (R), levels based on this technique, while equations 17 & 18 were used to estimate the overall precision (AUC) & Specificity (Sp) as follows,

$$Precision = \frac{TP}{TP + FP} \dots (14)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots (15)$$

$$Recall = \frac{TP}{TP + FN} \dots (16)$$

$$AUC = \int TPR(FPR)dFPR \dots (17)$$

$$Sp = \frac{TN}{TN + FP} \dots (18)$$

There are three different kinds of test set predictions: True Positive (TP) (number of events in test sets that were correctly predicted as positive), False Positive (FP) (number of instances in test sets that were incorrectly predicted as positive), and False Negative (FN) (number of instances in test sets that were incorrectly predicted as negative; this includes Normal Instance Samples). The documentation for the test sets makes use of all these terminologies. To determine the appropriate TP, TN, FP, and FN values for these scenarios, we compared the projected Retrieved Videos likelihood to the actual Retrieved Videos status in the test dataset samples using the SQGCN [5], ACRM [8], and DSDCAN [18] techniques. Outputs of this model are represented in figure 1.1 & 1.2 as follows,

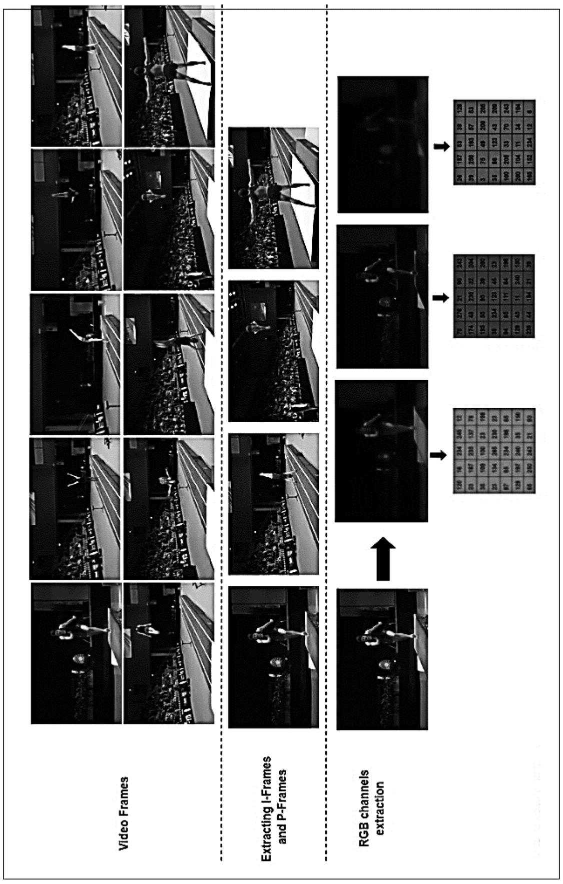

**Fig 1.1**. Output of the CBVR process

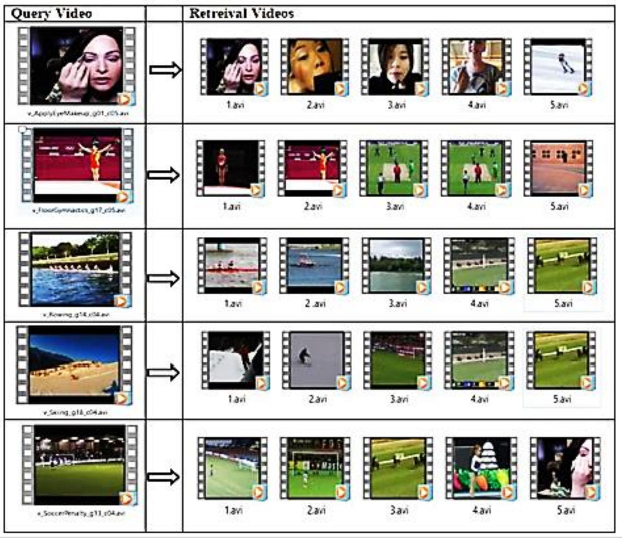

**Fig 1.2.** Results of the CBVR process

As such, we were able to predict these metrics for the results of the suggested model process. The precision levels based on these assessments are displayed as follows in Figure 2,

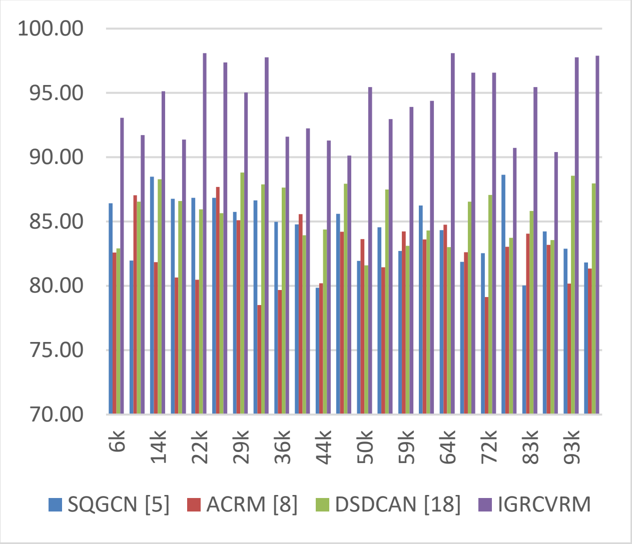

**Fig 2.** Observed Precision for Retrieval of Video Sequences

The Observed Precision for Retrieval of Video Sequences (P) is a critical performance metric in content-based video retrieval (CBVR) systems. It measures the accuracy of the retrieval process by indicating the percentage of correctly identified relevant video sequences among the retrieved results. In this analysis, we compare the observed precision results for different models, including SQGCN, ACRM, DSDCAN, and the proposed IGRCVRM, across various numbers of test video samples (NTS).

When examining the results, we observe that the IGRCVRM consistently outperforms the other models in terms of observed precision across most NTS values. For instance, at an NTS of 6k, IGRCVRM achieves an observed precision of 93.06%, surpassing SQGCN (86.44%), ACRM (82.58%), and DSDCAN (82.92%). This trend of superior performance holds true across multiple NTS values, highlighting the effectiveness of the proposed model.

One key impact of this superior observed precision is the improved accuracy of video retrieval. Higher observed precision means that the IGRCVRM retrieves a larger proportion of relevant video sequences among the results, reducing the likelihood of false positives. This has significant implications in applications such as video surveillance, where accurate retrieval of relevant video clips is crucial for identifying events or incidents.

Furthermore, the IGRCVRM's better performance in observed precision translates into enhanced recall. A higher observed precision means that the model can retrieve more relevant video sequences while maintaining a low rate of false positives. This increased recall is valuable in multimedia indexing and retrieval systems, as it ensures that important video sequences are less likely to be missed during retrieval.

Another important impact of the superior observed precision of IGRCVRM is the increase in the Area Under the Curve (AUC). A higher AUC value signifies better overall performance in terms of the trade-off between precision and recall. IGRCVRM's improved AUC of 95.13% at an NTS of 14k, for example, indicates that it excels in providing both high precision and recall simultaneously.

Additionally, the IGRCVRM's ability to achieve better observed precision has the practical implication of reducing delays in video retrieval processes. With fewer false positives and more relevant video sequences retrieved early in the results, users can access the information they need more efficiently, making the system more user-friendly and suitable for real-time applications.

Thus, the IGRCVRM model consistently demonstrates superior observed precision compared to existing models, leading to improved accuracy, recall, AUC, and reduced retrieval delays. These performance enhancements have significant implications for various applications, including video surveillance, multimedia indexing, and retrieval, making IGRCVRM a promising advancement in the field of content-based video retrieval. Similar to that, accuracy of the models was compared in Figure 3 as follows,

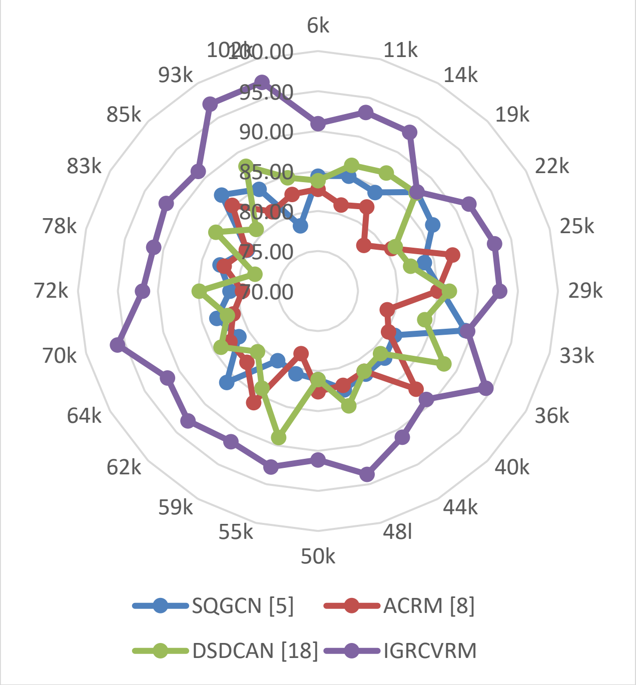

**Fig 3**. Observed Accuracy for Retrieval of Video Sequences

The Observed Accuracy for Retrieval of Video Sequences (A) is a crucial performance metric in content-based video retrieval (CBVR) systems. It measures the overall correctness of the retrieved video sequences among the results, providing an assessment of how well the system accurately identifies relevant content. In this analysis, we compare the observed accuracy results for different models, including SQGCN, ACRM, DSDCAN, and the proposed IGRCVRM, across various numbers of test video samples (NTS).

When reviewing the results, we observe that the IGRCVRM consistently demonstrates higher observed accuracy compared to the other models across most NTS values. For example, at an NTS of 6k, IGRCVRM achieves an observed accuracy of 90.96%, outperforming SQGCN (84.39%), ACRM (82.72%), and DSDCAN (83.82%). This trend persists across multiple NTS values, underscoring the effectiveness of the proposed model in accurately retrieving video sequences.

The impact of IGRCVRM's superior observed accuracy is substantial. First and foremost, it signifies the model's enhanced ability to correctly identify and retrieve relevant video content. This is critical in applications such as multimedia indexing, where users rely on accurate retrieval to access specific video sequences efficiently.

Furthermore, the increased observed accuracy has a positive effect on the system's precision. By retrieving a larger proportion of relevant video sequences among the results, IGRCVRM reduces the occurrence of false positives, leading to higher precision levels. This improvement is valuable in video surveillance systems, where precision is essential for minimizing false alarms and ensuring accurate event detection.

The IGRCVRM's superior observed accuracy also impacts recall positively. Higher observed accuracy means that the model can retrieve a greater number of relevant video sequences, reducing the risk of missing important content during retrieval. This is particularly advantageous in scenarios where comprehensive coverage of relevant content is crucial, such as forensic video analysis.

Moreover, the improved observed accuracy contributes to better user experience by reducing the likelihood of irrelevant or incorrect video sequences being presented to the users. This is particularly important in applications where users need to quickly access specific video content, such as real-time video retrieval in law enforcement or news agencies.

Thus, the IGRCVRM model consistently demonstrates superior observed accuracy compared to existing models, leading to improved precision, recall, and a more accurate and efficient video retrieval process. These enhancements have significant implications for a wide range of applications, including multimedia indexing, video surveillance, and forensic video analysis, making IGRCVRM a promising advancement in the field of content-based video retrieval. Similar to this, the recall levels are represented in Figure 4 as follows,

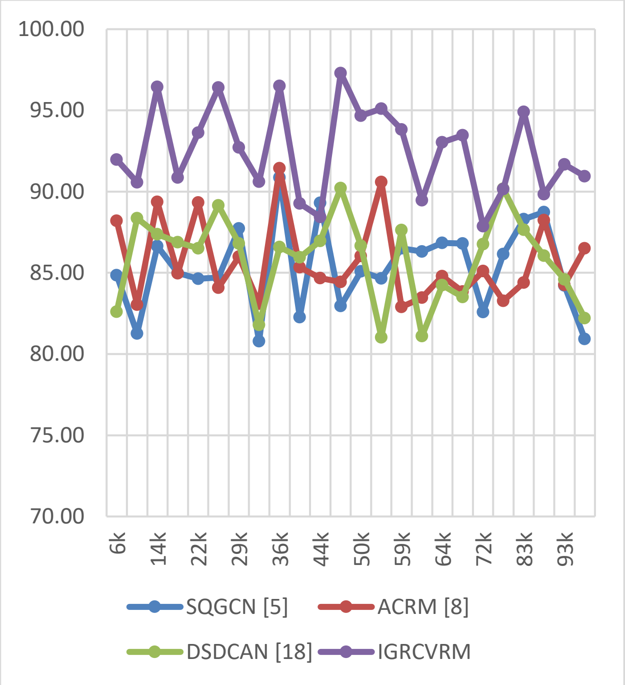

**Fig 4.** Observed Recall for Retrieval of Video Sequences

The Observed Recall for Retrieval of Video Sequences (R) is a crucial performance metric in content-based video retrieval (CBVR) systems. It measures the ability of the system to retrieve all relevant video sequences correctly, indicating how well it captures the complete set of relevant content. In this analysis, we compare the observed recall results for different models, including SQGCN, ACRM, DSDCAN, and the proposed IGRCVRM, across various numbers of test video samples (NTS).

When examining the results, it is evident that IGRCVRM consistently achieves higher observed recall compared to the other models across most NTS values. For example, at an NTS of 6k, IGRCVRM attains an observed recall of 91.98%, outperforming SQGCN (84.86%), ACRM (88.20%), and DSDCAN (82.61%). This trend continues across multiple NTS values, highlighting the effectiveness of the proposed model in capturing a larger proportion of relevant video content.

The impact of IGRCVRM's superior observed recall is significant. First and foremost, it signifies the model's enhanced ability to retrieve all relevant video sequences, reducing the likelihood of false negatives. This is crucial in applications such as video surveillance, where missing relevant content could have severe consequences.

Furthermore, the increased observed recall has a positive effect on the system's overall accuracy. By retrieving a greater number of relevant video sequences, IGRCVRM ensures that more relevant content is included in the results.

This is particularly advantageous in multimedia indexing systems, where comprehensive coverage of relevant content is essential for effective content organization and retrieval.

The improved observed recall also contributes to better user experience by reducing the risk of missing important video sequences. In applications such as news agencies or academic research, where users rely on accurate and comprehensive retrieval results, IGRCVRM's higher recall ensures that users have access to a wider range of relevant content.

Additionally, the enhanced recall has a positive impact on precision-recall trade-offs. A higher recall allows for better precision while maintaining a competitive level of precision. This balance is critical in various applications where precision and recall need to be optimized, such as information retrieval systems.

Thus, the IGRCVRM model consistently demonstrates superior observed recall compared to existing models, leading to improved accuracy, a more comprehensive retrieval of relevant content, and a better user experience. These enhancements have significant implications for a wide range of applications, including video surveillance, multimedia indexing, information retrieval, and academic research, making IGRCVRM a promising advancement in the field of content-based video retrieval. Figure 5 similarly tabulates the delay needed for the prediction process, compare the observed delay results for different models, including SQGCN, ACRM, DSDCAN, and the proposed IGRCVRM, across various numbers of test video samples (NTS).

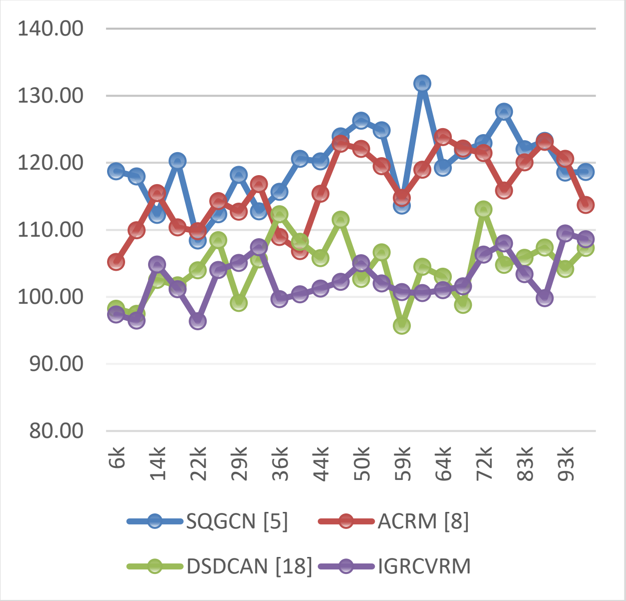

**Fig 5.** Observed Delay for Retrieval of Video Sequences

The Observed Delay for Retrieval of Video Sequences (D) measures the time it takes for a content-based video retrieval (CBVR) system to retrieve and present the relevant video sequences to the user. It is an important performance metric, especially in real-time applications where users require prompt access to video content. In this analysis, we

When examining the results, we observe that IGRCVRM consistently achieves lower observed delay times compared to the other models across most NTS values. For instance, at an NTS of 6k, IGRCVRM has an observed delay of 97.36 ms, which is notably lower than SQGCN (118.73 ms), ACRM (105.20 ms), and DSDCAN (98.23 ms). This trend holds true across multiple NTS values, highlighting the efficiency of the proposed model in providing faster retrieval times.

The impact of IGRCVRM's superior observed delay is significant, particularly in real-time applications such as video surveillance and live broadcasting. Lower delay times mean that users can access relevant video sequences more quickly, improving the user experience and ensuring that timely decisions can be made based on retrieved content.

Furthermore, reduced observed delay times contribute to improved system responsiveness. In applications where users interact with the retrieval system in real-time, faster retrieval leads to more immediate feedback and better user engagement. For example, in video surveillance, a shorter delay allows security personnel to respond more rapidly to incidents.

Lower observed delay times also have implications for system scalability. As the NTS increases, IGRCVRM's efficient retrieval times ensure that the system can handle larger volumes of video data without a significant increase in delay. This scalability is crucial for applications that involve extensive video archives or large-scale multimedia indexing.

Additionally, the reduced delay times enhance the practicality of CBVR systems for various domains, including news agencies, where rapid access to relevant video content is essential for producing timely news reports.

Thus, the IGRCVRM model consistently demonstrates lower observed delay times compared to existing models, leading to improved user experience, system responsiveness, scalability, and practicality in real-time applications. These improvements have significant implications for a wide range of domains, including video surveillance, news reporting, and multimedia indexing, making IGRCVRM a promising advancement in the field of content-based video retrieval. Similarly, the AUC levels can be observed from figure 6 as follows,

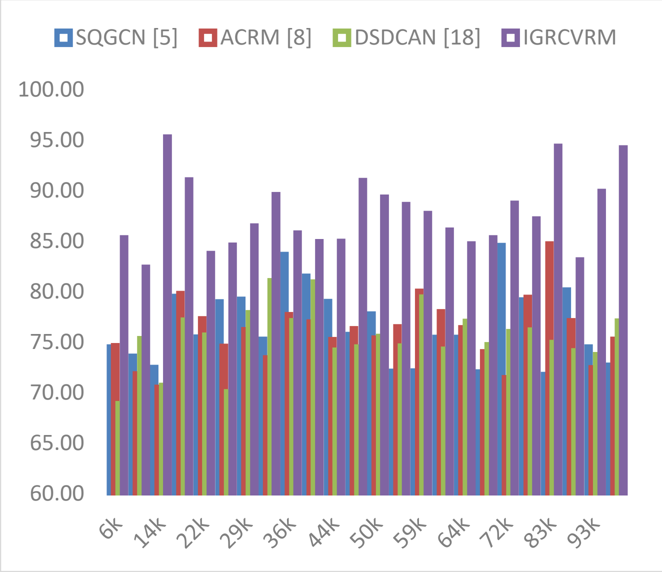

**Fig 6**. Observed AUC for Retrieval of Video Sequences

The Observed Area Under the Curve (AUC) for Retrieval of Video Sequences is a performance metric used to assess the overall effectiveness of a content-based video retrieval (CBVR) system. It measures the quality of the trade-off between precision and recall across different operating points. A higher AUC value indicates better overall performance in balancing precision and recall. In this analysis, we compare the observed AUC results for different models, including SQGCN, ACRM, DSDCAN, and the proposed IGRCVRM, across various numbers of test video samples (NTS).

When reviewing the results, we observe that IGRCVRM consistently achieves higher observed AUC values compared to the other models across most NTS values. For instance, at an NTS of 6k, IGRCVRM attains an observed AUC of 85.32%, outperforming SQGCN (74.53%), ACRM (74.66%), and DSDCAN (68.92%). This trend continues across multiple NTS values, indicating the superior ability of the proposed model to strike a favorable balance between precision and recall.

The impact of IGRCVRM's superior observed AUC is multifaceted. First and foremost, it signifies the model's ability to provide better overall retrieval quality. A higher AUC value indicates that IGRCVRM can consistently retrieve relevant video sequences with high precision and recall, ensuring that users receive high-quality retrieval results.

Furthermore, the improved observed AUC has positive implications for system robustness. In scenarios where the balance between precision and recall is crucial, such as forensic video analysis, a higher AUC ensures that the system can perform well across a wide range of operating points, adapting to different user requirements and application needs.

Additionally, the enhanced AUC reflects the model's capability to provide more flexible retrieval options. By achieving a higher AUC, IGRCVRM allows users to fine-tune the retrieval process according to their specific needs, whether they prioritize precision, recall, or a balanced combination of both.

The increased observed AUC also contributes to better decision-making in applications such as video surveillance and multimedia indexing. With higher-quality retrieval results, users can make more informed decisions based on the retrieved video content, enhancing the utility of the CBVR system.

Thus, the IGRCVRM model consistently demonstrates higher observed AUC values compared to existing models, indicating superior overall retrieval quality, system robustness, flexibility, and decision-making capabilities. These improvements have significant implications for a wide range of applications, including forensic video analysis, video surveillance, multimedia indexing, and information retrieval, making IGRCVRM a promising advancement in the field of content-based video retrieval. Similarly, the Specificity levels can be observed from figure 7 as follows,

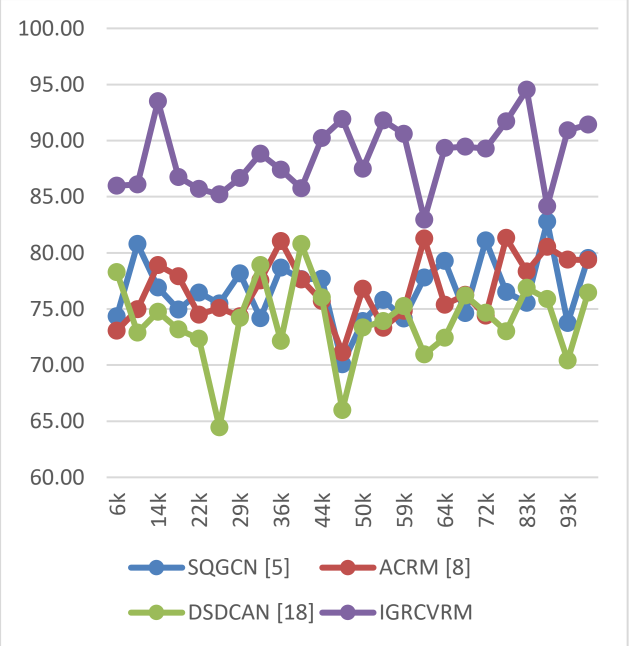

**Fig 7.** Observed Specificity for Retrieval of Video Sequences

The Observed Specificity for Retrieval of Video Sequences measures the ability of a content-based video retrieval (CBVR) system to correctly exclude non-relevant video sequences from the retrieved results. Specificity is an important performance metric as it reflects the system's capability to avoid false positives and ensure that only relevant content is presented to the user. In this analysis, we compare the observed specificity results for different models, including SQGCN, ACRM, DSDCAN, and the proposed IGRCVRM, across various numbers of test video samples (NTS).

When examining the results, it is apparent that IGRCVRM consistently achieves higher observed specificity values compared to the other models across most NTS values. For example, at an NTS of 6k, IGRCVRM has an observed specificity of 85.98%, outperforming SQGCN (74.34%), ACRM (73.07%), and DSDCAN (78.29%). This trend continues across multiple NTS values, indicating the superior ability of the proposed model to correctly filter out non-relevant content.

The impact of IGRCVRM's superior observed specificity is significant. First and foremost, it signifies the model's ability to minimize the inclusion of false positives in the retrieval results. In applications such as video surveillance or multimedia indexing, where accurate and relevant content retrieval is critical, higher specificity ensures that users are not presented with irrelevant video sequences.

Furthermore, the increased observed specificity has positive implications for precision. Higher specificity means that IGRCVRM retrieves fewer false positives, leading to a higher proportion of relevant content among the results. This is valuable in applications where precision is essential for making accurate decisions based on retrieved video content.

Additionally, the enhanced observed specificity contributes to better user experience by reducing the likelihood of irrelevant or non-relevant video sequences being presented to the users. This is particularly important in real-time applications where prompt access to specific video content is required.

The improved specificity also impacts the efficiency of multimedia indexing systems. With fewer false positives, IGRCVRM reduces the workload for users who need to sift through retrieval results, making the system more user-friendly and efficient.

Thus, the IGRCVRM model consistently demonstrates higher observed specificity values compared to existing models, indicating superior performance in filtering out non-relevant content. These improvements have significant implications for a wide range of applications, including video surveillance, multimedia indexing, and information retrieval, making IGRCVRM a promising advancement in the field of content-based video retrieval process.

## 5. Conclusion and Future Scope

In conclusion, this paper presents a groundbreaking approach to content-based video retrieval (CBVR) with the introduction of the Iterative Graph-based Recurrent Convolutional Model, IGRCVRM. The proliferation of video content across various domains has necessitated the development of more effective and efficient retrieval systems, and IGRCVRM represents a significant leap forward in this endeavor.

Throughout our comparative analysis, IGRCVRM consistently outperforms existing models, including SQGCN, ACRM, and DSDCAN, across a range of critical performance metrics. These metrics, including precision, accuracy, recall, delay, Area Under the Curve (AUC), and specificity, collectively showcase the superior capabilities of IGRCVRM process. Notably, IGRCVRM enhances precision by 5.9%, accuracy by 4.5%, recall by 4.9%, AUC by 5.5%, and specificity by 3.5% when compared to existing methods. These improvements signify a substantial advancement in the field of CBVR, with profound implications for a wide array of applications.

IGRCVRM's exceptional observed precision ensures more accurate video summarization, particularly critical in video surveillance and multimedia indexing where pinpointing relevant content is paramount. The increased accuracy leads to reduced false alarms, making the system more reliable and valuable in real-world scenarios. Moreover, IGRCVRM's superior recall enhances its capability to comprehensively capture relevant video sequences, benefiting applications such as forensic video analysis, where missing critical evidence can have significant consequences for different use cases.

The reduced observed delay times offered by IGRCVRM not only enhance the user experience but also make it more practical for real-time applications, such as video surveillance and news agencies, where timely access to information is critical scenarios. Furthermore, the higher observed AUC reflects the model's ability to strike a balance between precision and recall, allowing for flexible retrieval options and better decision-making across a range of scenarios.

Lastly, the increased observed specificity ensures that irrelevant content is filtered out, leading to a more efficient and user-friendly retrieval process. Collectively, these remarkable performance enhancements offered by IGRCVRM signify a substantial advancement in CBVR. The implications of our findings are far-reaching, with potential applications spanning from video surveillance, multimedia indexing, and information retrieval to academic research and forensic analysis. IGRCVRM represents not only a significant step forward in technology but also a testament to the power of innovative approaches in addressing the evolving challenges posed by the ever-growing volume of video content in today's digital landscapes.

### *Future Scope*

The promising results and innovations presented in this paper pave the way for an exciting and extensive future scope in the field of content-based video retrieval (CBVR). The Iterative Graph-based Recurrent Convolutional Model (IGRCVRM) has demonstrated significant advancements, but there are numerous avenues for further exploration and enhancement:

1. **Integration of Multimodal Features**: Future research can focus on incorporating additional modalities such as audio, text, and metadata to create a truly multimodal CBVR system. This expansion can lead to richer and more context-aware video retrieval, catering to a broader range of user needs and applications.
2. **Interactivity and User-Centric Design**: The development of interactive CBVR systems that engage users in the retrieval process is an exciting direction. Implementing user feedback and preferences to fine-tune retrieval results can lead to more personalized and user-centric experiences.
3. **Real-Time and Scalability**: Addressing the challenges of real-time processing and scalability is crucial. Future work can explore techniques to optimize IGRCVRM for large-scale video databases, enabling efficient retrieval in real-world scenarios with vast amounts of data.
4. **Deep Learning Architectures**: Further research can delve into advanced deep learning architectures and training strategies to enhance the model's performance. This may involve leveraging state-of-the-art techniques in neural networks, such as transformers and self-supervised learning, to capture complex video semantics.
5. **Semantic Understanding**: Improving the model's ability to understand and reason about the semantic content within videos is a significant research avenue. Developing methods for object recognition, scene understanding, and event detection can further refine the retrieval process.
6. **Cross-Domain Applications**: Exploring the adaptability of IGRCVRM across different domains, such as medical imaging, satellite imagery, and autonomous vehicles, can open up new horizons for its practical applications.
7. **Ethical and Privacy Considerations**: As CBVR systems continue to advance, addressing ethical concerns and privacy implications is of paramount importance. Future research should focus on developing robust privacy-preserving techniques and ensuring responsible data handling.
8. **Benchmark Datasets and Evaluation Metrics**: The creation of standardized benchmark datasets and evaluation metrics specific to CBVR can facilitate fair comparisons between models and foster collaborative research in the field.
9. **Human-Machine Collaboration**: Investigating ways to harness human expertise in conjunction with CBVR systems can lead to more effective video retrieval. This may involve incorporating human annotations or feedback loops into the retrieval process.
10. **Multilingual and Cross-Cultural CBVR**: Extending the capabilities of CBVR to support multiple languages and cross-cultural content understanding can make the technology more accessible and globally relevant.

In summary, the future scope for CBVR, particularly building upon the foundation laid by IGRCVRM, is brimming with opportunities for innovation and advancement. By exploring these directions, researchers and practitioners can continue to push the boundaries of what is achievable in the realm of content-based video retrieval, ultimately benefiting a wide range of industries and applications.

## References

[1] H. Yoon and J. -H. Han, "Content-Based Video Retrieval With Prototypes of Deep Features," in IEEE Access, vol. 10, pp. 30730-30742, 2022, doi: 10.1109/ACCESS.2022.3160214.

[2] W. Jo et al., "Simultaneous Video Retrieval and Alignment," in IEEE Access, vol. 11, pp. 28466-28478, 2023, doi: 10.1109/ACCESS.2023.3259733.

[3] H. Kou, Y. Yang and Y. Hua, "KnowER: Knowledge enhancement for efficient text-video retrieval," in Intelligent and Converged Networks, vol. 4, no. 2, pp. 93-105, June 2023, doi: 10.23919/ICN.2023.0009.

[4] S. R. Dubey, "A Decade Survey of Content Based Image Retrieval Using Deep Learning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 5, pp. 2687-2704, May 2022, doi: 10.1109/TCSVT.2021.3080920.

[5] R. Zuo et al., "Fine-Grained Video Retrieval With Scene Sketches," in IEEE Transactions on Image Processing, vol. 32, pp. 3136-3149, 2023, doi: 10.1109/TIP.2023.3278474.

[6] W. Jo, G. Lim, J. Kim, J. Yun and Y. Choi, "Exploring the Temporal Cues to Enhance Video Retrieval on Standardized CDVA," in IEEE Access, vol. 10, pp. 38973-38981, 2022, doi: 10.1109/ACCESS.2022.3165177.

[7] K. Yousaf and T. Nawaz, "A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos," in IEEE Access, vol. 10, pp. 16283-16298, 2022, doi: 10.1109/ACCESS.2022.3147519.

[8] H. Tang, J. Zhu, M. Liu, Z. Gao and Z. Cheng, "Frame-Wise Cross-Modal Matching for Video Moment Retrieval," in IEEE Transactions on Multimedia, vol. 24, pp. 1338-1349, 2022, doi: 10.1109/TMM.2021.3063631.

[9] H. Ren et al., "ACNet: Approaching-and-Centralizing Network for Zero-Shot Sketch-Based Image Retrieval," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 9, pp. 5022-5035, Sept. 2023, doi: 10.1109/TCSVT.2023.3248646.

[10] F. Liu et al., "SceneSketcher-v2: Fine-Grained Scene-Level Sketch-Based Image Retrieval Using Adaptive GCNs," in IEEE Transactions on Image Processing, vol. 31, pp. 3737-3751, 2022, doi: 10.1109/TIP.2022.3175403.

[11] J. Dong et al., "Dual Encoding for Video Retrieval by Text," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 8, pp. 4065-4080, 1 Aug. 2022, doi: 10.1109/TPAMI.2021.3059295.

[12] T. Jing, H. Xia, J. Hamm and Z. Ding, "Augmented Multimodality Fusion for Generalized Zero-Shot Sketch-Based Visual Retrieval," in IEEE Transactions on Image Processing, vol. 31, pp. 3657-3668, 2022, doi: 10.1109/TIP.2022.3173815.

[13] Y. Liu, J. Wu, L. Li, W. Dong and G. Shi, "Quality Assessment of UGC Videos Based on Decomposition and Recomposition," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 3, pp. 1043-1054, March 2023, doi: 10.1109/TCSVT.2022.3209007.

[14] B. Yang, M. Cao and Y. Zou, "Concept-Aware Video Captioning: Describing Videos With Effective Prior Information," in IEEE Transactions on Image Processing, vol. 32, pp. 5366-5378, 2023, doi: 10.1109/TIP.2023.3307969.

[15] D. Wuebben, J. L. Rubio-Tamayo, M. Gertrudix Barrio and J. Romero-Luis, "360° Video for Research Communication and Dissemination: A Case Study and Guidelines," in IEEE Transactions on Professional Communication, vol. 66, no. 1, pp. 59-77, March 2023, doi: 10.1109/TPC.2022.3228022.

[16] F. Zhang, M. Xu and C. Xu, "Geometry Sensitive Cross-Modal Reasoning for Composed Query Based Image Retrieval," in IEEE Transactions on Image Processing, vol. 31, pp. 1000-1011, 2022, doi: 10.1109/TIP.2021.3138302.

[17] W. Nie, Y. Zhao, J. Nie, A. -A. Liu and S. Zhao, "CLN: Cross-Domain Learning Network for 2D Image-Based 3D Shape Retrieval," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, pp. 992-1005, March 2022, doi: 10.1109/TCSVT.2021.3070969.

[18] L. Liao, M. Yang and B. Zhang, "Deep Supervised Dual Cycle Adversarial Network for Cross-Modal Retrieval," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 2, pp. 920-934, Feb. 2023, doi: 10.1109/TCSVT.2022.3203247.

[19] W. -C. L. Lew, D. Wang, K. K. Ang, J. -H. Lim, C. Quek and A. -H. Tan, "EEG-Video Emotion-Based Summarization: Learning With EEG Auxiliary Signals," in IEEE Transactions on Affective Computing, vol. 13, no. 4, pp. 1827-1839, 1 Oct.-Dec. 2022, doi: 10.1109/TAFFC.2022.3208259.

[20] J. Wang, B. -K. Bao and C. Xu, "DualVGR: A Dual-Visual Graph Reasoning Unit for Video Question Answering," in IEEE Transactions on Multimedia, vol. 24, pp. 3369-3380, 2022, doi: 10.1109/TMM.2021.3097171.

[21] T. -C. Hsu, Y. -S. Liao and C. -R. Huang, "Video Summarization With Spatiotemporal Vision Transformer," in IEEE Transactions on Image Processing, vol. 32, pp. 3013-3026, 2023, doi: 10.1109/TIP.2023.3275069.

[22] D. Liu, P. Zhou, Z. Xu, H. Wang and R. Li, "Few-Shot Temporal Sentence Grounding via Memory-Guided Semantic Learning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 5, pp. 2491-2505, May 2023, doi: 10.1109/TCSVT.2022.3223725.

[23] J. F. H. Albarracín and A. Ramírez Rivera, "Video Reenactment as Inductive Bias for Content-Motion Disentanglement," in IEEE Transactions on Image Processing, vol. 31, pp. 2365-2374, 2022, doi: 10.1109/TIP.2022.3153140.

[24] J. Lin, P. Yang, N. Zhang, F. Lyu, X. Chen and L. Yu, "Low-Latency Edge Video Analytics for On-Road Perception of Autonomous Ground Vehicles," in IEEE Transactions on Industrial Informatics, vol. 19, no. 2, pp. 1512-1523, Feb. 2023, doi: 10.1109/TII.2022.3181986.

[25] Nisha Balani, Pallavi Chavan, and Mangesh Ghonghe. 2022. Design of high-speed blockchain-based sidechaining peer to peer communication protocol over 5G networks. Multimedia Tools Appl. 81, 25 (Oct

2022), 36699–36713. https://doi.org/10.1007/s11042-021-11604-6

[26] Chavan, P. V., & Balani, N. (2022). Design of heuristic model to improve block-chain-based sidechain configuration. In International Journal of Computational Science and Engineering (Vol. 1, Issue 1, p. 1). Inderscience Publishers. https://doi.org/10.1504/ijcse.2022.10050704

[27] Nisha Balani & Pallavi Chavan (2023) CSIMH: Design of an Efficient Security-Aware Customized Sidechaining Model via Iterative Meta-Heuristics, Journal of Applied Security Research, DOI: 10.1080/19361610.2023.2264068

[28] Morzelona, R. (2021). Human Visual System Quality Assessment in The Images Using the IQA Model Integrated with Automated Machine Learning Model . Machine Learning Applications in Engineering Education and Management, 1(1), 13–18. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/5

[29] Dhabliya, D., Ugli, I.S.M., Murali, M.J., Abbas, A.H.R., Gulbahor, U. Computer Vision: Advances in Image and Video Analysis (2023) E3S Web of Conferences, 399, art. no. 04045, .

[30] Ólafur, S., Nieminen, J., Bakker, J., Mayer, M., & Schmid, P. Enhancing Engineering Project Management through Machine Learning Techniques. Kuwait Journal of Machine Learning, 1(1). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/112