Intelligent Decision Technologies 17 (2023) 687–712
DOI 10.3233/IDT-220303
IOS Press

Enhancement of video condensation with video browsing and retrieval for surveillance videos using heuristic-derived deep learning algorithm

Suhandas[a,*] and Santhosh Kumar G[b]
[a]*Department of Electronics and Communication Engineering, A.J. Institute of Engineering and Technology, Kottara Chowki, Mangaluru, India*
[b]*Department of Electronics and Communication Engineering, East West College of Engineering, Yelahanka New Town, Bengaluru, India*

Abstract. Video condensation or synopsis is an effective solution for problems regarding video storage and video browsing. The proposed model contributed to developing the video condensation framework for efficient video browsing and video retrieval. In the first stage, the videos are gathered from the surveillance videos. Here, the frames are generated, and then the video backgrounds are extracted. The objects from the frames are acquired through the support of Yolov3. Next, the optimal stitching is done based on the time and object activity of video frames using the Improved Blue Monkey Optimization (IBMO) algorithm. Moreover, video condensation is performed to get the compact video for making better browsing and retrieval of video. The video browsing and retrieval are performed under two phases such as training and testing phases and both phases are done by gathering the videos and followed by the feature extraction using VGG16, where the heuristic improvement is made by the same IBMO algorithm. Then, the extracted deep features from video segments are clustered based on Fuzzy C-means (FCM) clustering for combining the extracted features. These features are stored in the feature database in the training phase. Next, in the testing phase, video browsing and retrieval are performed by considering the queries gathered from the standard dataset. The features of query videos are extracted, which are compared based on Multi-Similarity Function (MSF) with the features in the database for retrieving the video segments. Experimental results show that the developed IBMO-VGG-MSF-based video condensation saves computational loads compared to the previous methods without compromising the condensation ratio and visual quality.

Keywords: Video condensation, video browsing and retrieval, surveillance videos, improved blue monkey optimization, multi-similarity function, optimal stitching, VGG16, fuzzy C-means clustering

1. Introduction

Humans have implemented various manual video retrieval schemes, which are more tedious and time-consuming. Moreover, these manual systems are highly prone to errors and produce incorrect retrieval results. The current technological era needs automatic schemes for the retrieval of videos concerning offline videos and internet videos to support users' specifications [1]. The video retrieval systems are applicable in diverse areas of data science that include education, entertainment, video archiving, surveillance, news, advertising, and also over the medical domain. In large cities, human activities are monitored and recorded

*Corresponding author: Suhandas, Department of Electronics and Communication Engineering, A.J. Institute of Engineering and Technology, Kottara Chowki, Mangaluru 575006, India. E-mail: suhandas099@gmail.com.

ISSN 1872-4981/$35.00

by surveillance cameras, which have been fixed widely in outdoor and indoor environments [2]. A large set of surveillance cameras are adopted for recording all events 24 hours/a day. The specific targets are searched by watching surveillance videos, which are highly sensitive to labor-intensive [3]. The important task is to effectively utilize the surveillance videos and obtain valuable information from the recorded surveillance videos. This may generate more time for retrieving, so the video condensation methods are developed. Here, the relevant segments are retrieved from the original surveillance videos based on the user queries, and the video condensation methods can acquire specific targets very quickly [4]. The retrieval of exact segments from large surveillance videos based on semantic queries is a challenging issue nowadays [5]. Several minutes of browsing on a several-hour video extraction are more helpful for the users since it assures minimum time for video retrieval. The main aim of this developed browsing system is to filter only the relevant segments by preserving all other object activities to get the specific target from the long duration of surveillance videos [6], which provides better support to the users. Many researchers considered the video retrieval systems in the good literature reviews that have been published. The challenging task is understanding the semantic gap between video content and user intentions [7].

The surveillance video contains a wide range of behavior, and hence, the retrieval of videos based on the understanding ability of the user requests and user query is very difficult [8]. This difficulty arises since there is no clear idea about the specific targets for making the video retrieval system. In addition, another challenge is to extract the moving objects from the original surveillance video and segment the moving objects related to the specific condition under a real-world environment [9]. The factors such as occlusion, shading, and lighting may affect the segmentation results. In usual scenarios, unexpected factors make it difficult to retrieve moving and tracing objects [10]. The collection of all compact frames is condensed and preserved based on the location and also gets customized for the target objects on the background images [11]. The three main characteristics of the compact video are that (a) it is short enough for video browsing; (b) it is considered as the significant component related to moving objects in the source video, and (iii) it preserves the temporal features of moving objects partially for the users to understand the original video. The system directly displays the appropriate video clip by localizing the targets rapidly when the user can browse the compact video [12]. The main aim of this video condensation and video retrieval system is to spend a short period of time watching and monitoring long surveillance videos.

Deep learning achieves greater success in video retrieval in several applications like visual tracking, object recognition, and visual analysis, where the deep features are directly learned from the pixels and voxels in the videos [13]. Moreover, the objects with background and moving targets are effectively identified based on this deep learning structure [14]. The features for video representation are efficiently extracted using Deep learning networks like Convolutional Neural Networks (CNN), and the temporal features are eliminated with the support of hashing function [15]. The manual detection process needs more time to retrieve the objects from the image or video, which can be resolved by deep learning techniques. The hidden factors are learned automatically by using multi-level nonlinear mappings in the neural networks [16]. The accuracy of the video retrieval process is highly increased using these approaches [17]. Hence, this research focused on developing a new video browsing and retrieval system using a deep learning structure.

The major contribution of the developed deep learning-based video browsing and retrieval system is described below.

- To design an efficient deep learning-aided video condensation model for video browsing and retrieval system to quickly retrieve video clips or contents from long-duration videos based on the user queries in surveillance video.
- To develop an effective video condensation system, where the background is extracted from the long video initially, and then the moving objects are detected using the YOLOv3 classifier. Frame stitching is carried out, where the stitched frames are optimally selected using the developed IBMO based on

the objects' activity and time interval to get the condensed video with a highly minimized object uncovered rate.
- To implement an IBMO for selecting the frames to be stitched in the condensed video and optimizing the parameters such as epochs and the steps per epochs in the VGG16 network to enhance the F1-score in the retrieval system.
- To develop a video browsing and retrieval process in a surveillance system based on user queries, where the features are extracted from the condensed video using VGG16 and then generated the clustered segments using FCM, and then MSF is used to check the similarity of the query and stored data to appropriately retrieve the respective videos.
- The efficiency of the developed model is validated over various heuristic algorithms and existing video retrieval models in terms of precision, recall, and F1-score.

The remaining section used in the developed video browsing and video retrieval system is given as follows. Section 2 describes the existing video retrieval systems with their features and challenges. Section 3 summarizes the dataset details and the description of the structural model. Section 4 explains YOLOv3-based object detection, the proposed IBMO algorithm, and video condensation. Section 5 illustrates the VGG16-based feature extraction, FCM-based clustering, and testing and training phase with user queries. Sections 6 and 7 provide the experiment results and conclusion of the developed video browsing and retrieval system.

2. Literature survey

2.1. Related works

In 2015, Chieh and Fang [18] have recommended a surveillance video browsing and retrieval system for locating the desired targets quickly to the appropriate users. The most significant information regarding moving objects was gathered from surveillance videos for constructing the relevant compact video. The compactness of the video was increased by rearranging the temporal coordinates of the moving objects from the relevant compact video. The essential activities from the original surveillance video have been preserved based on the visual appearance of moving objects. The watching time of the compact video has been highly minimized by using the developed model by avoiding the monitoring of long surveillance videos for several hours. Moreover, several experiments have been conducted to demonstrate the quick look at specific targets from the surveillance videos via the newly implemented system.

In 2017, Ding et al. [19] have designed a large surveillance video retrieval system by exploiting the data characteristics and big data processing procedures. The entire system has been functioning based on the motion information from the videos. Initially, the video was segmented concerning the relevant data, and then the basic unit was named M-clip after segmentation. The data volume was highly reduced by neglecting the redundant video content via the M-clips. The human detection and extraction of motion or appearance features have been done via the MapReduce framework for processing the M-clips. Only the sub-areas instead of entire frames have been processed through vision algorithms on the significant motion vectors. The developed model outperformed by evaluating the experimental results based on satisfactory human retrieval accuracy with computational time.

In 2020, Cheng et al. [20] have proposed a deep learning and cloud-based face video retrieval system to provide greater accuracy. Initially, the gathered data was pre-processed to remove the blurs and then performed the face alignment on the remaining images. After that, pre-training was done via FaceNet, ArcFace, and VGGFace for face recognition. The final results have been compared to three different models to choose the most efficient one to develop the system. Moreover, the system's feasibility has been

verified by the implementation of the prototype in the proposed system. The implementation outcome ensured that the developed system performed well than the other models concerning computational time and recognition accuracy.

In 2020, Poornima and Saleena [21] have presented a deep learning strategy that adopted a video retrieval scheme, where keyframe extraction was performed for the retrieval of useful keyframes from the original video. The retrieved keyframe features were stored in the feature database. For instance, the FCM procedure has been utilized to cluster the retrieved features. Consequently, these clustered features have been given into the deep structure to determine the optimal centroid. Different categories of videos have been considered for the experimentation for performing the retrieval based on both the video query and the text query. From the test results, the developed video retrieval system attained improved performance than other video retrieval systems based on the consideration of certain measures like F-measure, recall, and precision.

In 2022, Kumar and Seetharaman [22] have offered a Modified Visual Geometry Group _16-based deep learning strategy for the extraction of features from the video. The indexing value has been assigned to all video files to improve the efficiency of the video retrieval process. The experimental results were compared among various video extraction approaches such as Convolution Neural Networks (CNN), Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG). The developed system has provided elevated video retrieval performance when analyzing measures like precision, F1 score, recall, and accuracy.

In 2020, Ullah et al. [23] have introduced a pre-trained 3D-CNN-based event-oriented feature selection framework by deeply investigating the response and weights to a particular event. Here, the neurons were semantically eliminated from the original video, which did not respond to an event. Additional storage was needed for storing the event-oriented convolutional features because they have huge dimensions and require more time to retrieve features. Then, the Principle Component Analysis (PCA) was utilized for generating compact binary codes from these features. The major benefit of this developed video retrieval system was to provide very efficient results over large-scale databases. The implementation results have been verified over HMDB51 and UCF101 datasets, and the created compact codes accomplished greater effectiveness according to the recall, precision, and execution time.

In 2020, Nguyen et al. [24] offered a video condensation scheme to fast monitoring of moving objects in surveillance videos, which has a long duration for the entire video. The effectiveness of the video condensation algorithm has been analyzed based on the condensation ratio and computational complexity. At first, the non-moving objects from the video frames were discarded. Secondly, frame grouping has been done via the intra condensation over the moving objects and then done inter- condensation. The temporal static pixels and the spatiotemporal static pixels among Intra and inter-condensation were dropped to shorten the temporal distances. The effectiveness of the developed model was observed to be high when analyzing the experimental results of this video condensation without sacrificing the visual quality and condensation with a less computational load than the conventional video condensation methods.

In 2015, Zhu et al. [25] have proposed an online content-aware framework for providing a video condensation system. The tube rearrangement of the optimization problem has been converted into a stepwise optimization problem in this video condensation system. Hence, this developed model has achieved a higher convergence rate and less memory requirement when compared to the offline framework. The condensed videos have been obtained simultaneously by using this transformation technique, and the proposed system has been highly suitable for real-time endless surveillance videos. The experimental result has been analyzed over various video condensation schemes, and the results were shown that the developed model achieved higher performance in terms of execution speed.

2.2. Problem statement

Some recent video retrieval techniques are reviewed in Table 1. Temporal sequencing [18] is more helpful in discovering the specific targets for users in surveillance videos and increases the compactness of

the video by minimizing the processing time. It does not apply to the GPU implementation of our system. MapReduce framework [19] achieves better human retrieval accuracy and takes less computational time even processing large-scale surveillance videos. It suffers from breakages in video semantics during the video segmentation process. CNN [20] has reached higher recognition accuracy on the face from the videos and reduces the computational time, and increases the feasibility of the system. The real-time design of this framework is limited. DBN [21] efficiently retrieves the data regarding F-measure, precision, and recall and achieves superior performance because of the enhanced key frame extraction process. It suffers from a high processing time when retrieving text and video queries. CNN [22] shows better performance on video frame retrieval regarding metrics like F1 score, precision, recall, and accuracy and achieves better performance with the help of extracting the video image frames. It is not applicable for processing larger datasets. 3D-CNN [23] effectively retrieves videos from huge-scale databases and reduces the execution time, and attains superiority in terms of recall and precision. It suffers from processing small-scale datasets. Intra-GoFM [24] achieves faster and order-preserving condensation and reduces the computational burden, and enhances performance. It suffers from processing long sequential videos. SILTP [25] attains better video condensation performance and achieves higher outcomes regarding condensed video with lower memory. It suffers from background consistency. Thus, there is a need to suggest a video condensation model for effective video browsing and retrieval, especially for surveillance applications. The advantages of the offered video retrieving system are listed below. While retrieving the text and video queries, the designed method provides better performance and it provides a low processing time. Moreover, it can perform in real-time large complex datasets to provide the enhanced performance of the designed model. Hence, it provides a clear background consistency from the retrieved videos. Consequently, it provides sufficient results to enhance the reliability of the system's performance.

3. A novel framework for surveillance video condensation with video browsing and retrieval using deep learning concept

3.1. Video condensation with video browsing and retrieval framework

For security and monitoring purposes, video surveillance systems are adopted in many venues, such as shopping malls, railway stations, and airports. Searching for people in crowded places is very challenging with long-duration surveillance videos based on feature representations due to the changes in weather and climate conditions. Rapid retrieval of moving objects in surveillance videos is necessary and desirable in a broad spectrum of real-world applications. Watching surveillance videos for searching specific targets with the support of humans are highly labor intensive, and extracting valuable information from long-duration videos is also very difficult. Several techniques are designed for retrieving video segments from a specific period to provide proper content corresponding to the user requirement, which is observed to be a very difficult task in video surveillance systems. The important challenge is that the skipped frames have missed some necessary content from the video. Several video condensation algorithms are developed based on extracting spatial and temporal features from the video, which gives a low condensation ratio, but if the nearest objects are in different directions, then the condensation performance is very poor. Hence, several tube filling-based condensation algorithms are developed to provide efficient results, but they suffer from the requirement of excessive memory storage for storing the videos. To address these issues, a deep structure-based video condensation and video browsing and retrieval system are developed by adopting heuristic algorithms to provide quick results over user-specified videos from large-duration videos. The structural representation of the developed video browsing and retrieval system is given in Fig. 1.

An efficient video condensation adopted video browsing and video retrieval system are designed, which helps users quickly look into the target of interest that is highly required over the long surveillance videos.

Table 1
Features and challenges of existing deep learning-based video retrieval systems

Author [citation]	Methodology	Features	Challenges
Chiehand Fang [18]	Temporal sequencing	– This method is more helpful for users to discover the specific targets in surveillance videos. – It increases the compactness of the video by minimizing the processing time.	It does not apply to GPU implementation of our system.
Ding et al. [19]	MapReduce framework	– It achieves better human retrieval accuracy. – It takes less computational time even to process large-scale surveillance videos.	It suffers from a high processing time with the retrieval of text and video queries.
Cheng et al. [20]	CNN	– It has reached higher recognition accuracy on the face from the videos. – It reduces the computational time and increases the feasibility of the system.	The real-time design of this framework is limited.
Poornima and Saleena [21]	DBN	– It efficiently retrieves the data regarding F-measure, precision, and recall. – It achieves superior performance because of the enhanced key frame extraction process.	It suffers a high processing time owing to retrieving text and video queries.
Kumar and Seetharaman [22]	CNN	– It performs better on video frame retrieval regarding metrics like F1 score, precision, recall, and accuracy. – It is better with the help of extracting the video image frames.	It is not applicable for processing larger datasets.
Ullah et al. [23]	3D-CNN	– It effectively retrieves the videos from a huge scale database. – It reduces the execution time and attains superiority regarding recall and precision.	It suffers from processing small-scale datasets.
Nguyen et al. [24]	Intra-GoFM	– It achieves faster and order-preserving condensation. – It reduces the computational burden and enhances performance.	It suffers from processing long sequential videos.
Zhu et al. [25]	SILTP	– It attains better video condensation performance. – This model achieves higher outcomes regarding condensed video with lower memory.	It suffers from background consistency.

The basic idea is to gather all the objects that, include the background and the moving objects, which carry the most relevant information in the surveillance videos and then construct a single compact video based on the extracted information. Firstly, the required large surveillance videos are taken from publicly available databases, and the background is extracted from the videos. The background is extracted in the format of keyframes. Secondly, the moving objects from the videos are extracted using the YOLOv3 model based on the time and object activity in the video. Consequently, the frame stitching is carried out over the detected objects, where the frames to be stitched are determined with the help of the developed IBMO algorithm. Thus, the selected frames with detected objects are stitched in the background frame, and finally, the condensed video is obtained. The objective function of the developed IBMO algorithm is to minimize the rate of objects uncovered from the surveillance videos. Thirdly, the condensed video is given to video browsing and retrieval operation, where the condensed video is segmented into many frames, and the deep features are extracted using the VGG16 network. In this VGG16, the epochs and the steps per epoch are optimized using developed IBMO to maximize the F1-score. These extracted deep features are clustered into more segments using the FCM clustering algorithm to get segmented videos, and then the clustered segments are stored in the feature database. This storing process is done in the training phase, and the testing phase, user queries (image or text) are stored in the database, and the features are extracted from the queries, which are subjected to MSF concerning Euclidean distance and cosine similarity for comparison of previously stored clustered video segments. If the Euclidean distance and cosine similarity

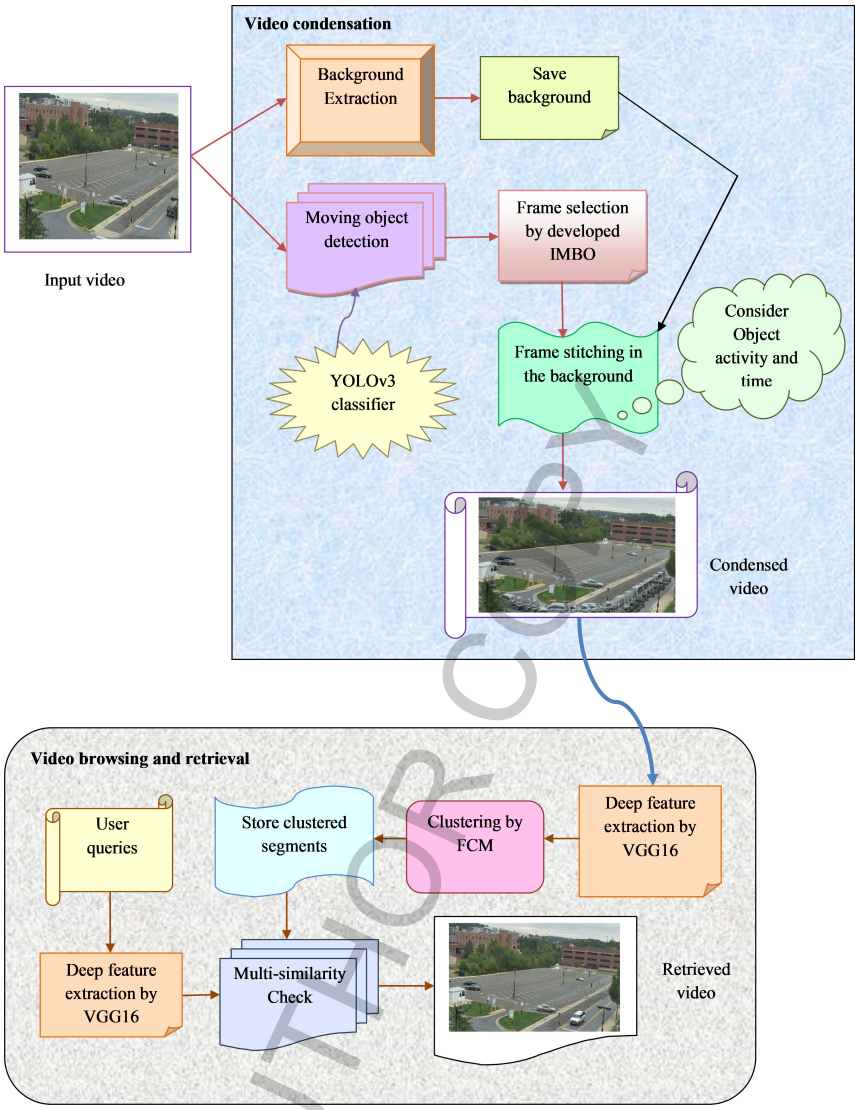

Fig. 1. Architectural illustration of developed deep structure-based video browsing and retrieval system.

values are observed to be low, then the retrieval of the appropriate video from the database occurs based on a user query within a few minutes. The efficiency of the developed video browsing and video retrieval system is analyzed with various conventional models regarding precision, recall, and f1-score.

3.2. Surveillance videos dataset collection

The surveillance videos are obtained from the Kitware database, and the name of dataset 1 is 'VIRAT' which is available on the online source of "https://data.kitware.com/#collection/56f56db28d777f753209ba9f/folder/56f581ce8d777f753209ca43" (access date 2022-22-11). The video files are available in three types of datasets that are public dataset, VIRAT video dataset Release 2.0, and the VIRAT ground dataset. The public dataset consists of annotations, documents, homographic, software, and original videos. VIRAT video dataset Release 2.0 contains VIRAT Aerial Dataset and VIRAT ground dataset. More video files are present in this dataset at various sizes.

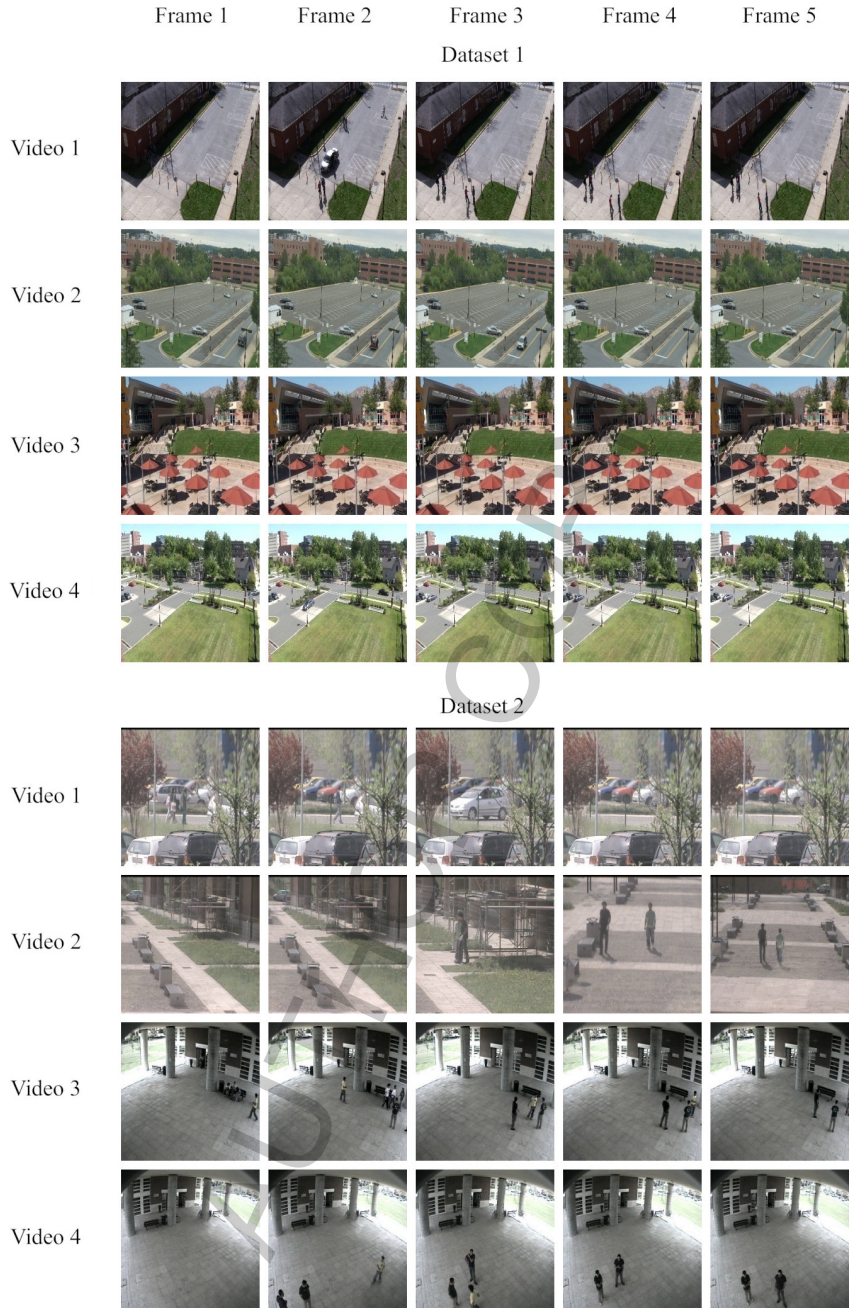

Fig. 2. Sample surveillance videos from Dataset 1 and Dataset 2.

Dataset 2 utilized for this video retrieval scheme is "Visor" which is available in the online source of "https://aimagelab.ing.unimore.it/visor/video_videosInCategory.asp?idcategory=11" (access date: 2022-25-11). It includes the video with base annotation, automatic annotations, GT annotations, reference papers, VidVideo corpus set, attachments, and other related videos in VISOR. For every individual video, base annotation details such as structural annotation and ground truth base annotation are described. Moreover, the details like video details, date of creation, type of camera, and infrared capabilities are also given. The sample surveillance videos are given in Fig. 2.

The input videos are indicated by VD_x^{Inp} where $x = 1, 2, 3, \ldots, X$ and X represent the total number of videos. At first, the background is extracted from the large surveillance videos and then constructs frames based on the extracted background. The extracted background is indicated as BG_n^{Inp}.

4. Optimal video condensation process using improved blue monkey optimization

4.1. YOLOv3-based object detection

The frames from the large surveillance videos VD_x^{Inp} are given as the input of the YOLOv3 model for detecting moving objects. YOLOv3 [26] is the deep learning framework used for object detection in smart surveillance systems. Generally, YOLOv3 is a Fully Convolutional Network (FCN), and the detection is obtained by employing the probability regression scores and the bounding boxes. In the developed video condensation system, the YOLOv3 is used to detect objects because; it is one of the fastest object detection models in real-time. The accuracy of the YOLOv3 network-based object detection is high when compared to other models. The mean average precision is also high in the YOLOv3 model. The Darknet-53 acts as the backbone of the YOLOv3 network, which performs the extraction of features using the input videos. The feature maps are influenced by the Feature Pyramid Network for extracting features. The backbone network contains a total of 53 CNN layers, followed by 3×3 and 1×1 FCN layers for enabling object detection. The prediction is precisely done at three scales by downsampling strides at 32, 16, and 8, respectively.

The detection kernel used by the YOLOv3 model is $1 * 1 * (M * (5 + N))$, where the number of probable classes is indicated by N, the bounding box relative number is denoted by M and the number of five bounding box offsets is presented. The fine-grained features from the smaller objects are effectively detected by the concatenation of unsampled layers. The primary blocks in the YOLOv3 object detector are residual box and skip connection, grid cells, bounding box regression, and Intersection over Union (IoU) block.

Residual box and skip connections: The residual layers are used in the YOLOv3 object detector to vanish the gradients by determining the deviations in the identity layers. The output obtained from one layer is subjected to an input of the next layer to avoid convergence degradation problems. Instead of performing direct mapping, skip connections are used in the YOLOv3 object detector to ensure the direct transfer of input from the immediate layer to the further layer.

Grid cells: The real-time video is divided into $P \times P$ dimension grids by the YOLOv3 object detector. The equal dimension grids are used to detect the object exactly from the real-time videos.

Bounding box regression: The objects are highlighted by providing outlines in the video frame with the help of bounding boxes. The output layer determines the confidence probability, dimensions, objectness score, and the coordinates of the bounding boxes.

IoU: The essential feature of the YOLOv3 object detector is the IoU, which helps to describe the overlapping of bounding boxes. It evaluates the similarity between the round truth bounding box and the predicting bounding box. Then, the comparison is made between the ratio of the overlapping region to the total combined area that is represented below Eq. (1).

$$IoU = \frac{Area\ to\ be\ overlapped}{Combined\ area} \tag{1}$$

Finally, object detection is done in the YOLOv3 with the help of IoU and bounding boxes. The detected objects are denoted by DO_k^{Inp}. The YOLOv3-based moving object detection is diagrammatically represented in Fig. 3.

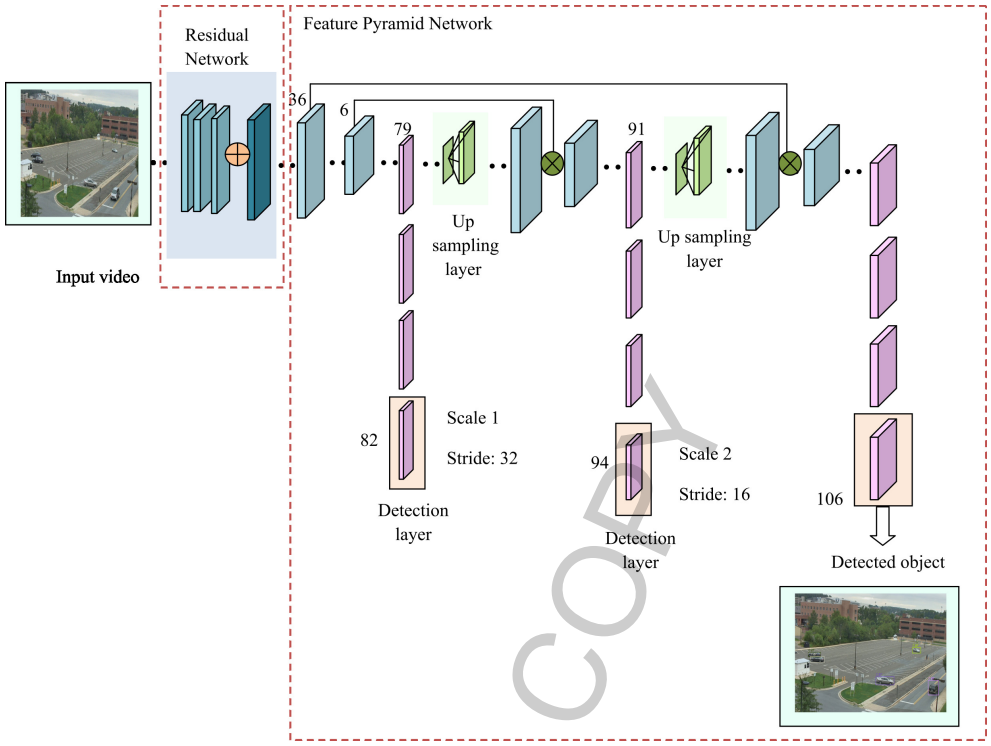

Fig. 3. Moving object detection using YOLOv3 from original videos.

4.2. *Proposed IBMO*

The developed IBMO is helpful for the selection of video frames that are to be stitched during video condensation to minimize the object uncovered rate, and it is also useful to optimize the parameters like epochs and steps per epochs in VGG16 during deep feature extraction to enhance the F1-score of resultant video retrieval results. This algorithm is selected in the developed video browsing and video retrieval system because of its flexibility, simplicity, and also avoiding local optima problems. But, it faces several problems like gradient and low probability during optima stagnation. Hence, improvement is needed in the BMO [27] to rectify these issues and thus, the IBMO is made. In the conventional algorithm, the parameter $r_1$in Eq. (4) is randomly selected but in the developed IBMO, it is considered using the new concept given in Eq. (2).

$$r_1 = r \tag{2}$$

Here, the parameter r represents the randomly created number chosen in the interval between [0, 1]. Similarly in Eq. (5), the parameter $r_2$is selected randomly in the conventional algorithm, but in the developed IBMO, it is calculated using the adaptive formula in Eq. (3).

$$r_2 = 1 - r \tag{3}$$

In Eq. (3), the parameter r denotes the randomly generated number in the interval between [0, 1]. The monkey's position is updated until reaching the best solution in the search space.

BMO: It has converged through the global optima with the support of optimization problems. The unidentified search space and the real-world problems with restrictions are effectively resolved through this algorithm. The inspiration and the mathematical model are briefly summarized below.

Algorithm 1: Recommended IBMO

```
Initialize the population of blue monkeys and children
Initialize the weight L and power rate R of blue monkeys and children
Distribution of blue monkeys and children into teams
Set the population size Po_S and maximum number of iterations MA_It
  For i = 1 to Po_S
  For j = 1 to MA_It
    Evaluate the fitness of blue monkeys and children in the group
    Set the parameter r
    Determine the value r_1 using Eq. (2)
    Determine the value r_2 using Eq. (3)
  For each group
    Find the best and worst value and store the best value as the current best
  End For
  End For
  While t ⩽ MA_It
    Swapping the worst fitness by the best fitness value in each group
    Update the power rate and position of the blue monkey using Eq. (4)
    Update the power rate and position of children using Eq. (6)
    Check whether the new best is better than the current best
    t = t + 1
  End For
  End while
Return the best optimal blue monkey
```

Inspiration: The BMO has been developed based on the behavior of the blue monkey. Based on the place searching for food at long distances, the monkeys are divided into teams, and the male mitis has no interaction with the younger ones. One male and more females with babies are presented in the divided groups.

Update position: The position updating depends on the position of the best blue monkey in the divided groups, and that position is delineated by the below Eq. (4).

$$R_{f+1} = (0.7 * R_f) + (L_{wt} - L_f) * r_1 * (M_{bst} - M_f) \tag{4}$$

Here, the power of the monkeys is indicated by R , and leader weight is represented by L_{wt}, M_f denotes the position of the monkey, and the term M_{bst} represents the position of the leader, weight of the monkey is defined by L_f, where all the weights are created randomly in the interval of [4, 6], and the term r_1 represents the arbitrary number in the range between [0, 1].

The updated position of the leader monkey is mathematically represented the below Eq. (5).

$$M_{f+1} = M_f + R_{f+1} * r_2 \tag{5}$$

The updated power rate of the monkey is indicated by the term R_{f+1}, M_f denotes the position of the monkey, and r_2 denotes the arbitrary number.

Then, the position is updated for the child blue monkey based on the monkey power rate and position of the previous iteration, which is shown in Eq. (6).

$$R_{f+1}^{CH} = (0.7 * R_f^{CH}) + (L_{wt}^{CH} - L_f^{CH}) * rand * (M_{bst}^{CH} - M_f^{CH}) \tag{6}$$

Here, the child power rate is represented by R^{CH}, child leader weight is denoted by L_{wt}^{CH}, child position is represented by L_f^{CH}, the term L_{wt}^{CH} indicates the leader child position, the child weight is indicated byL_f^{CH}, where all weights are selected in the interval of [4, 6], and the term *rand* is the arbitrary number in the range of [0, 1]. Here, the term M_{bst}^{CH} is defined as the position of the child. The variable M_f^{CH}represents the position of the leader child.

The updated position of the leader child monkey is given in Eq. (7).

$$M_{f+1}^{CH} = M_f^{CH} + R_{f+1}^{CH} * rand \tag{7}$$

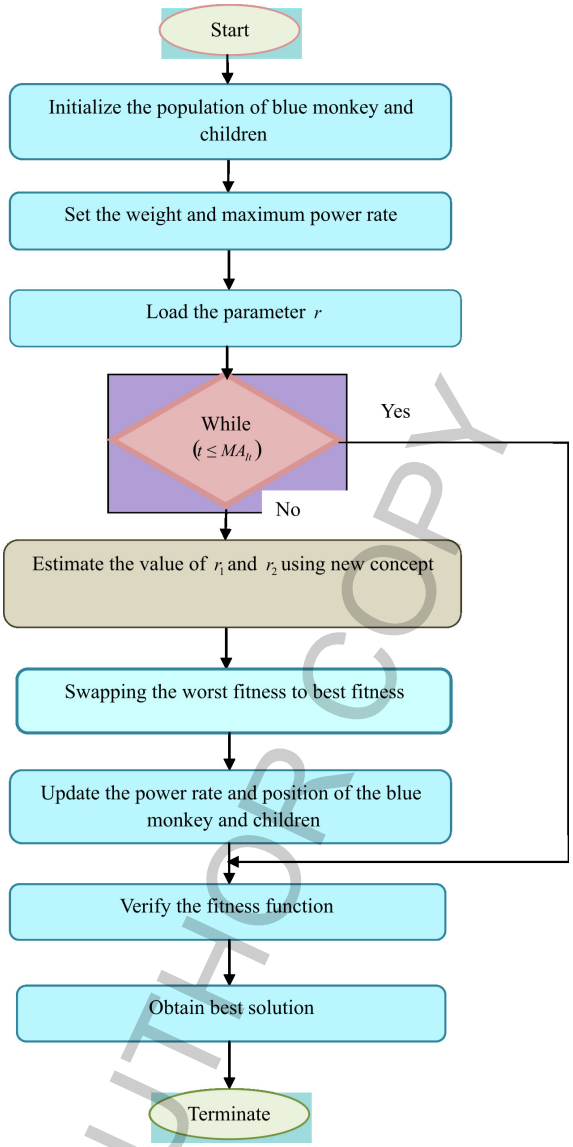

Fig. 4. Flowchart of the developed IBMO.

The updated power rate of the child monkey is indicated by M_{f+1}, the position of the child monkey is indicated by M_f, and the term *rand* gives the arbitrary number. The position of the monkeys is updated for all iterations. The pseudocode of the developed IBMO algorithm is enumerated in Algorithm 1.

The flowchart of the developed IBMO is given in Fig. 4.

4.3. Video condensation process by IBMO-based optimal stitching

Video condensation is mainly used for compressing long-duration videos. This video condensation is required to achieve quick and efficient retrieved videos based on the user's queries from the lengthy surveillance videos. This video condensation helps to reduce the inactive density of the videos to enhance the speed of the retrieval process. Initially, the keyframes from the background are extracted because the background frames are fixed. Then, the moving objects are extracted from the surveillance video,

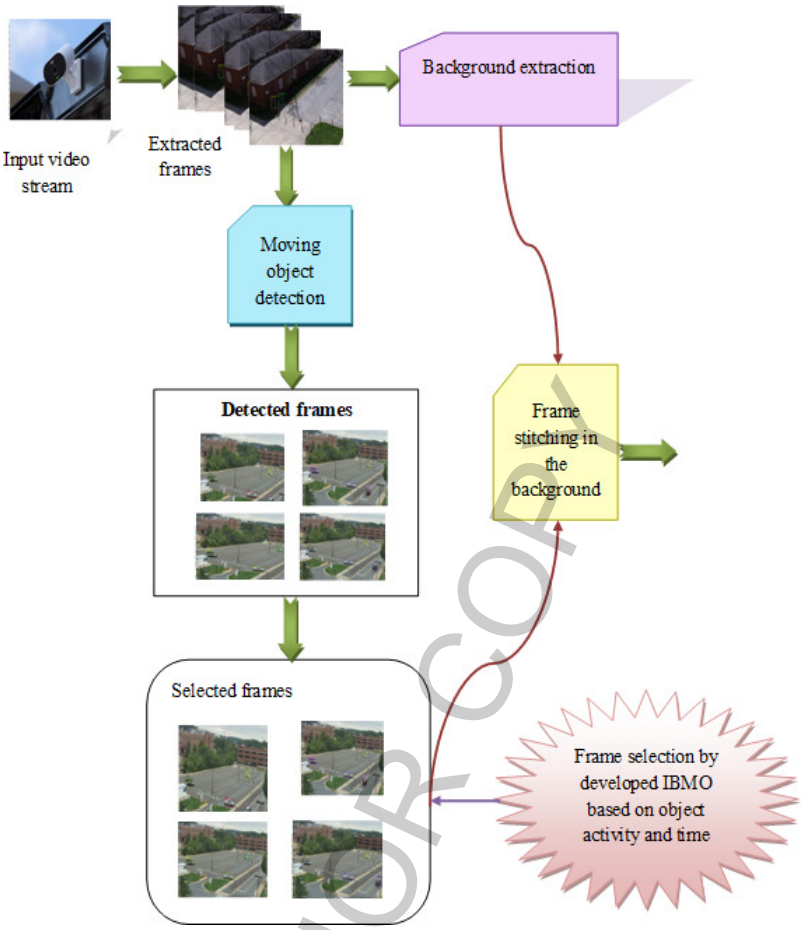

Fig. 5. Video condensation based on IBMO-based frame stitching.

and the patterns related to moving objects are noted and stitched into a single frame. If any unexpected changes in the background, then they are specifically saved. The extraction of moving objects is done by the Yolov3 classifier, and then the extracted moving objects are stitched based on the time interval and the activity of the objects. While stitching frames, the appropriate frames are selected using the developed IBMO algorithm that is indicated by the term FR_p^{CON}. Before stitching, the phase correlation related to the overlapping region is calculated, and then find the exact feature points are using developed IBMO for stitching the frames. Finally, matched regions are calculated for stitching, and the detected objects with selected frames are stitched in the background frame. The objective function of the developed video condensation algorithm is to reduce the objects uncovered in the videos. The objective function is given in below Eq. (8).

$$FJ_1 = \underset{\{FR_p^{CON}\}}{\arg\min}(OB(Not)_{cvr}) \tag{8}$$

Here, the term FR_p^{CON} defines the selected frames to be stitched via the developed IBMO, and the term $OB(Not_{cvr})$ gives the object to be uncovered from the video. The objects to be uncovered are estimated in Eq. (9).

$$OB(Not_{cvr}) = (OB_{TOT} - OB_{CON}) \tag{9}$$

Here, the term $OB(Not_{cvr})$ denotes the objects uncovered in the video, OB_{TOT} gives the total number of objects in the video, and the term OB_{CON} represents the objects in the condensed video. The objects uncovered are highly minimized, and then it seems that all the objects are condensed in a single video without any information loss, and the performance of the video condensation is high while minimizing this object's uncovered rate. Finally, all the objects are almost included in the single condensed video. The schematic representation of video condensation using IBMO-based optimal frame stitching is given below in Fig. 5.

5. New multi-similarity-based video browsing and retrieval using deep learning and optimization in surveillance system

5.1. Proposed video browsing and retrieval process

The compressed video is given to the video browsing and retrieval process. The deep structure-based video browsing and video retrieval system are used to provide appropriate videos based on user queries regarding specified times and objects from the long-duration surveillance videos. The time duration required to give appropriate videos is very less compared to other models. Initially, the compressed videos are segmented into separate frames based on the moving objects, and then the segmented frames are stored in the library. From these stored frames, the deep features are extracted using the VGG16 network, where the epochs and the steps per epoch are optimized using the developed IBMO for maximizing the F1 score of the video retrieval process. Then, the FCM clustering approach is used to get the clustered video segments that are stored in the feature database. Then, the user gives a query about the long-duration video with a particular object, and that query will be saved in the feature database. From these stored data, the desirable features are extracted, and the features are subjected to the multi-similarity check with the previously stored clustered data. If the multi-similarity score is high and then the video segments are retrieved from the clustered video segments. This video browsing and retrieval process are done via two phases like training phase and the testing phase. In the training phase, the user queries are stored, and features are extracted from the queries and in the testing phase, checking the multi-similarity score between the user queries and the already stored video segments. Similarity checking is carried out with the help of computing the Euclidean distance and cosine similarity. If the Euclidean distance and cosine similarity values are less, then the video related to those queries is retrieved from the database very quickly. The training and testing during video browsing and retrieval are given in Fig. 6.

5.2. Feature extraction by optimal VGG16

In the developed video browsing and retrieval system, the VGG16 network extracts features from the appropriate keyframes. The VGG16 [28] architecture consists of a maximum pooling layer, twelve convolutional layers, four fully connected layers, and finally, the softmax layer. The applied frames are transformed into different sizes before the extraction of features. In the 1st and 2nd convolution layer, it uses a total of 64 feature maps with the size of 3×3 and the stride of 14. Hence, the dimension of the frame is transformed 224×244 into $224 \times 244 \times 64$. Then, the maximum pooling is applied with a filter size of 3×3 and output stride of 2 and, further, obtains the frame with a dimension of $224 \times 244 \times 64$ into the dimension of $112 \times 112 \times 64$.

The reduced dimension frame is passed through the 3rd and 4th convolutional layer, which uses a 128 feature map with filter size 3×3 and output stride 2. The dimension of the image is further reduced $56 \times 56 \times 128$. Then, this frame is given to the 5th and 6th convolutional layers by considering 256 layers with

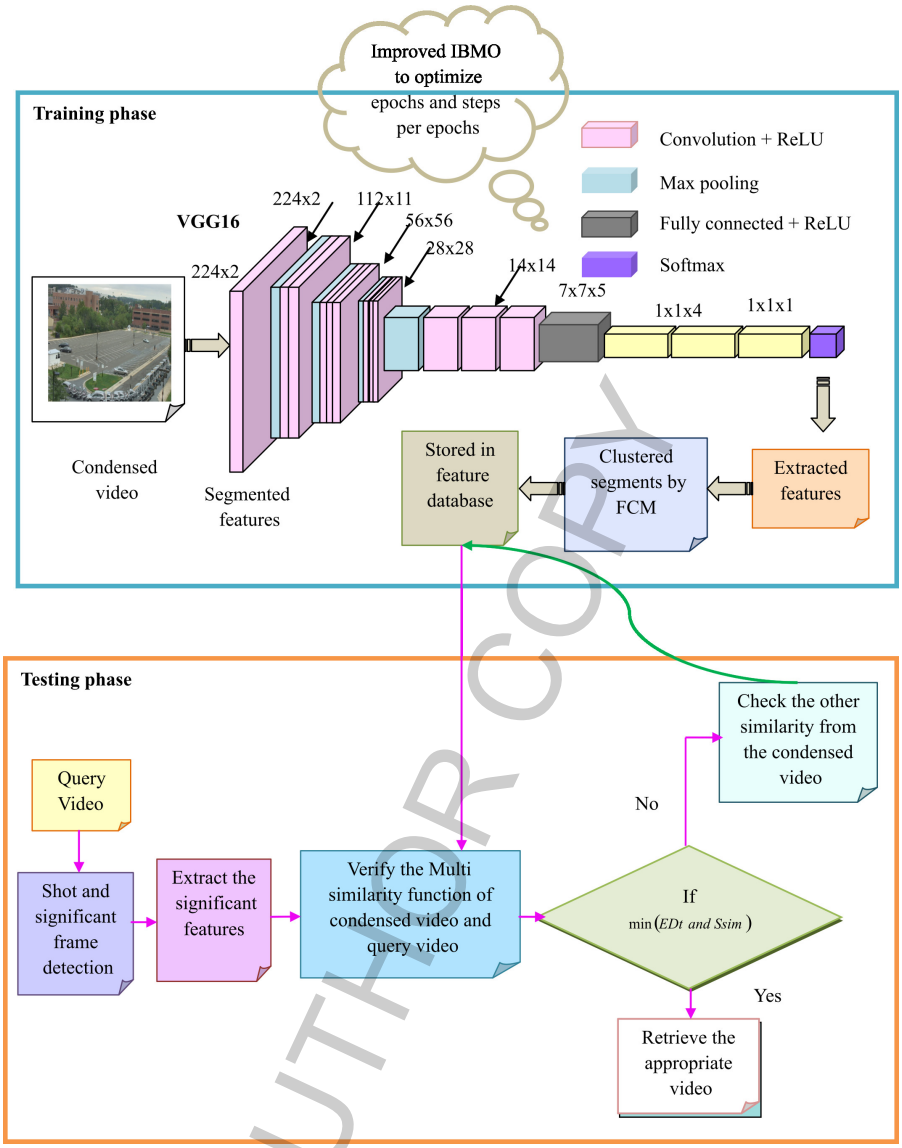

Fig. 6. Training and testing process in video browsing and retrieval model.

filter size 3 × 3 and output stride 1. The maximum pooling layer is applied concerning 256 feature maps with 3 × 3 filter size and output stride as 2.

Finally, two sets of convolutional layers are present inside the seventh to twelfth layer, which follows the maximum pooling layer that uses the 512 feature maps with filter size 3 × 3 with an output stride of 1 and then gets the frame with a dimension of 7 × 7 × 512. The most relevant features are extracted from the condensed videos using this VGG16 network. The extracted features are indicated by EX_t^{VGG16}. The parameters, such as epochs and steps per epoch, are optimized to increase the F1-score in VGG16. The objective function of this mechanism is to increase the F1-score that is mathematically expressed in Eq. (10).

$$FJ_2 = \underset{\{Eo_l^{VGG}, SP_m^{VGG}\}}{\arg\min} \left(\frac{1}{F1\text{-}Score} \right) \tag{10}$$

Here, the term Eo_l^{VGG} indicates the optimized epochs in the VGG16 and SP_m^{VGG} denotes the optimized step per epoch in the VGG16. Furthermore, the term *F1-Score* represents the obtained F1-score value of the video retrieved system. F1-score combines the recall and the precision of a classifier into a single metric and then calculates the harmonic mean between the recall and precision. It is evaluated in Eq. (11). From datasets 1 and 2, the optimized epochs were considered to evaluate better system performance. In dataset 1, the optimized epochs of [55, 118] are taken, and the optimized epochs of [62, 141] are considered for dataset 2. Moreover, the learning rate for datasets 1 and 2 is considered as -0.001.

$$F1\text{-}Score = \frac{2\alpha}{2\alpha + \beta + \delta} \tag{11}$$

The F1-score value mainly depends on the true positive and negative measures and the true negative and false negative measures. The term α represents the true positive observation value, β gives the true negative measure, and δ is the false negative measure.

5.3. FCM-based clustering

FCM clustering algorithm combines the deep features from the VGG16 model for segmenting the videos. The input given to the FCM is EX_t^{VGG16}. In fuzzy [29], let us consider that $B = \{k_s,\ s = 1, 2, \ldots, Q \mid k_s \in \mathrm{M}^p\}$ is the feature extracted frames with Q pixels. These frames are divided into w clusters. The term k_s represents the feature data, and the formation of clusters is given in Eq. (12).

$$P_y = \sum_{l=1}^{T}\sum_{s=1}^{U} \vartheta_{cls}^{y} \|k_s - \psi_l\|^2 \tag{12}$$

The constraints are represented in Eqs (13) and (14), respectively.

$$\sum_{l=1}^{T} \vartheta_{ls}^{y} = 1, \quad \forall s; \qquad 0 \leqslant \psi_{ls} \leqslant 1, \quad \forall l, s; \tag{13}$$

$$\sum_{s=1}^{U} \vartheta_{ls} > 0, \quad \forall l \tag{14}$$

Here, ψ_{ls} gives the membership function of the pixel k_s , and the class center value is indicated by the term ψ_l. The Euclidean distance between the pixel and weights is denoted by the term $\|\cdot\|$. Finally, the video segments are obtained and stored in the feature database.

5.4. Multi-similarity function derived for video retrieval

The use of this developed IBMO-VGG16-MSF-based video browsing and video retrieval system is to provide appropriate content based on the user queries from the large surveillance videos. The MSF is used in the testing phase for checking the similarity of the features of user queries with previously stored clustered features in the feature database. The metrics to be utilized for performing the MSF are Euclidean distance and Cosine similarity. The value of Euclidean distance and Cosine similarity needs to be low between the query features and features in the database for retrieving appropriate video segments from the database respective to the query. The MSF is given in the below Eq. (15).

$$MSF = Min\{(EDt) + (Csim)\} \tag{15}$$

Here, the term *EDt* represents the Euclidean distance, and the term *Csim* denotes the cosine similarity. Euclidean distance is the distance between the two videos, and it is given in Eq. (16).

$$EDt(A, B) = \sqrt{\sum_{n=1}^{p} (B_n - A_n)^2} \tag{16}$$

Here, the term A_n is the feature vectors from the database, and the term B_n is the feature vectors from the query video.

Cosine similarity is the ratio of the dot product of the feature vectors to the product of video lengths given in Eq. (17).

$$Csim = \frac{\sum_{n=1}^{p} A_n B_n}{\sqrt{\sum_{n=1}^{p} A_n^2}\sqrt{\sum_{n=1}^{p} B_n^2}} \tag{17}$$

If the Euclidean distance and Cosine similarity value are low, then the appropriate video is retrieved from the condensed video.

6. Results and discussions

6.1. Experimental setup

The newly suggested IBMO-VGG16-MSF-based video condensation with video browsing and retrieval system was implemented in Python v.3.9.13. The population size to be taken for conducting experiments was 10, the chromosome length was two, and the maximum number of iterations to be taken was 25 by comparing with various video browsing and retrieval systems. Recently developed video browsing and video retrieval models were taken to compare the performance of the developed model in terms of precision, recall, and F1-score. The previously used video browsing and video retrieval models like CNN [20], DBN [21], K-Means [30], and FCM [29]. Moreover, the heuristic algorithms to be considered for evaluating the effectiveness of the developed video browsing and video retrieval system were Artificial Gorilla Troops Optimization (AGTO) [31], Water Strider Algorithm (WSA) [32], Deer Hunting Optimization Algorithm (DHOA) [33] and BMO [27]. The implementation platform used in the research work is Python v.3.9.13. Moreover, the processor is an Intel Core i3 of 64-bit processor and the operating system of 64-bit in Windows 10. Hence, the RAM size is 16 GB.

6.2. Performance metrics

Precision, recall, and F1-score are the performance measures used to analyze the effectiveness of various video browsing and video retrieval algorithms. The formula for Precision, recall, and F1-score are explained as follows.

Precision: Precision is a good measure, which is computed by dividing the number of correct results by the number of all the returned results. It is mathematically defined in Eq. (18).

$$Precision = \frac{\alpha}{\alpha + \gamma} \tag{18}$$

Here, the term γ indicates the false positive observation value α is the true positive measure.

Recall: Recall is determined by dividing the number of true negatives by the total number of elements in the positive class. It is given in Eq. (19).

$$Recall = \frac{\alpha}{\alpha + \delta} \tag{19}$$

F1-score: F1-score is calculated using Eq. (11).

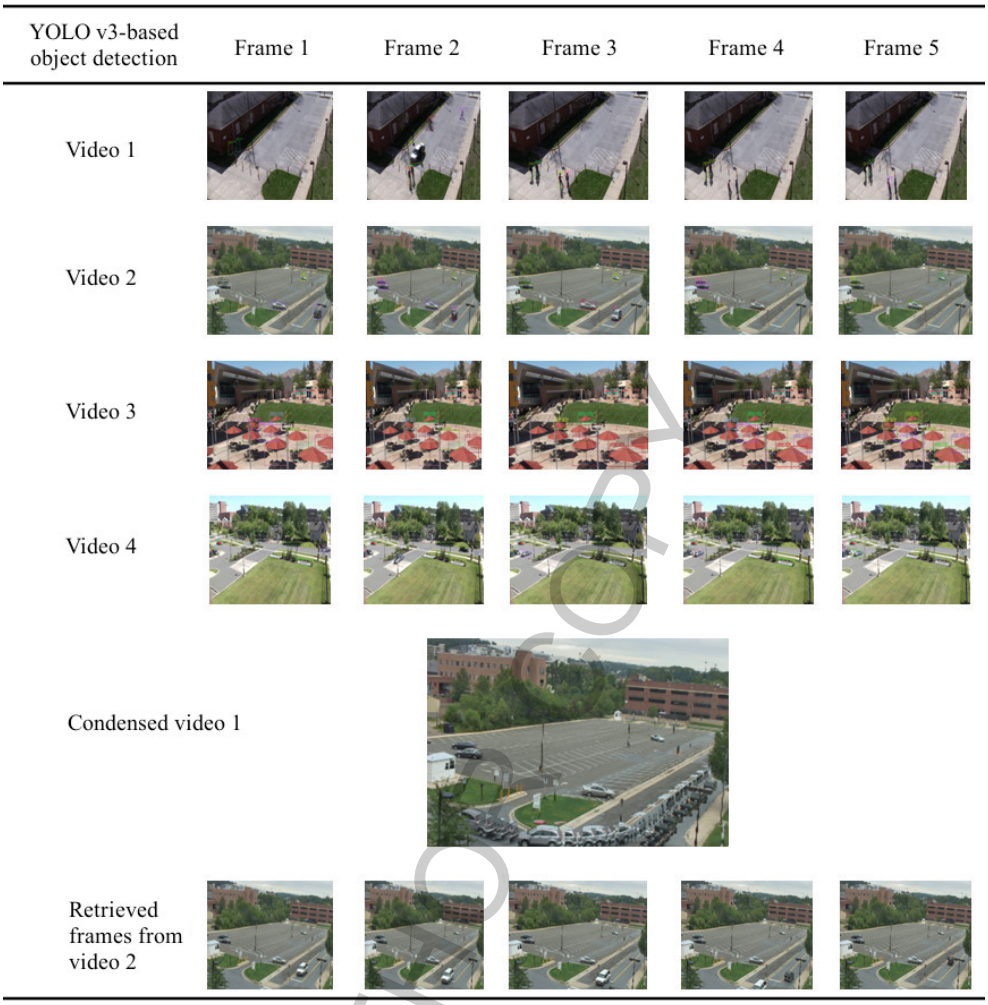

Fig. 7. Resultant retrieved frames from the condensed video.

6.3. *Experimental results*

The results obtained from the developed video browsing and retrieval system are given in Fig. 7.

6.4. *Performance analysis using dataset 1*

The performance comparison of the developed IBMO-VGG16-MFO-based video browsing and video retrieval system over different heuristic algorithms and distinct existing video retrieval systems using dataset 1 is correspondingly depicted in Figs 8 and 9. The precision, recall, and F1-score measures are taken for the analysis according to the number of received frames. From this plot, the developed video retrieval system accomplished an improved recall of 12.65% than BMO, 13.37% than WSA, 14.10% than AGTO, and 15.58% than DHOA according to the received frames at 10. In addition, from Fig. 9, the developed video retrieval system performed better than CNN, DBN, K-means, and FCM regarding the performance measures like precision, recall, and F1-score.

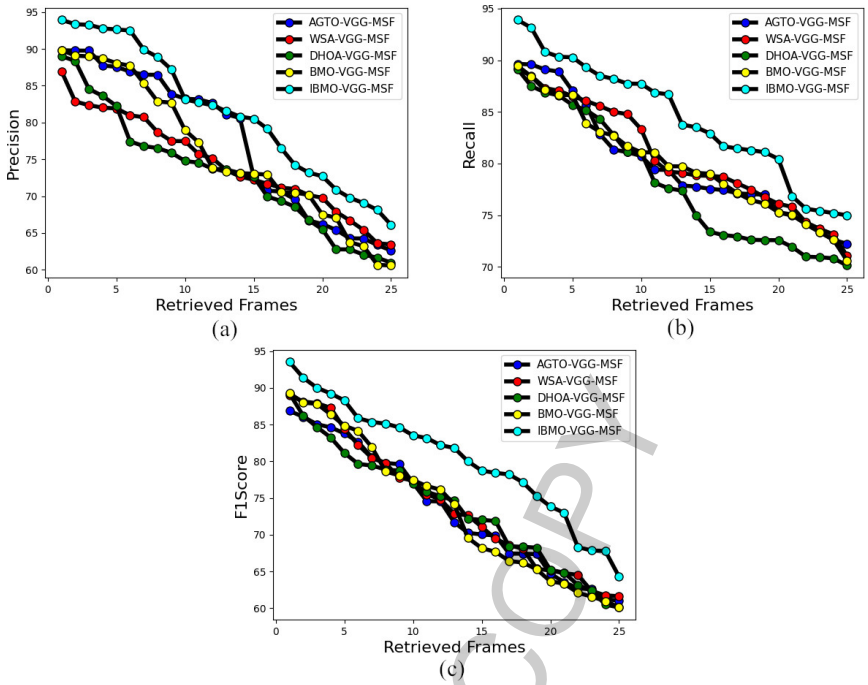

Fig. 8. Comparing the efficiency of the developed video retrieval system on dataset 1 among different heuristic strategies in regards to (a) Precision (b) Recall and (c) F1-score.

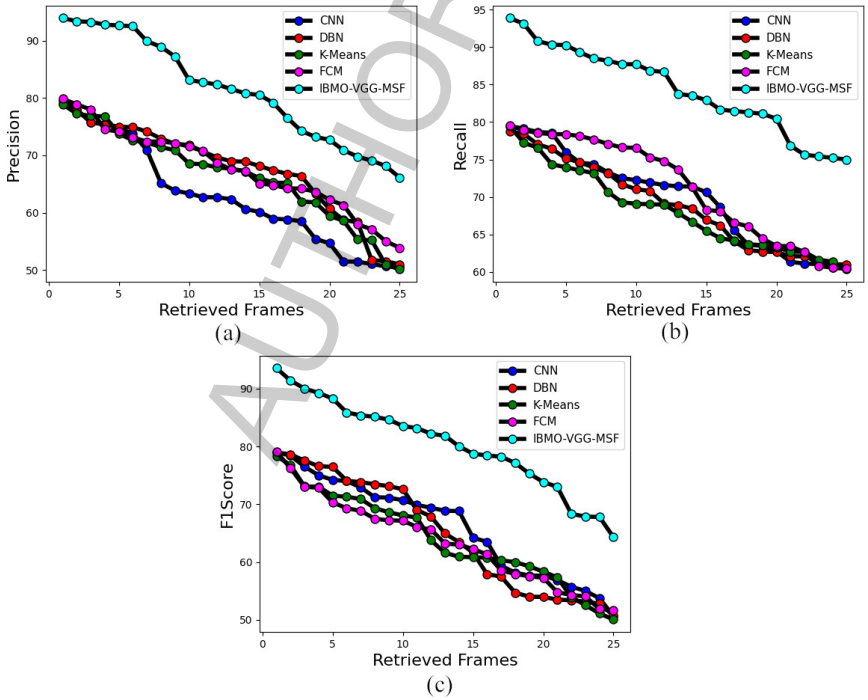

Fig. 9. Comparing the efficiency of the developed video retrieval system on dataset 1 among conventional video retrieval techniques in regards to (a) Precision, (b) Recall, and (c) F1-score.

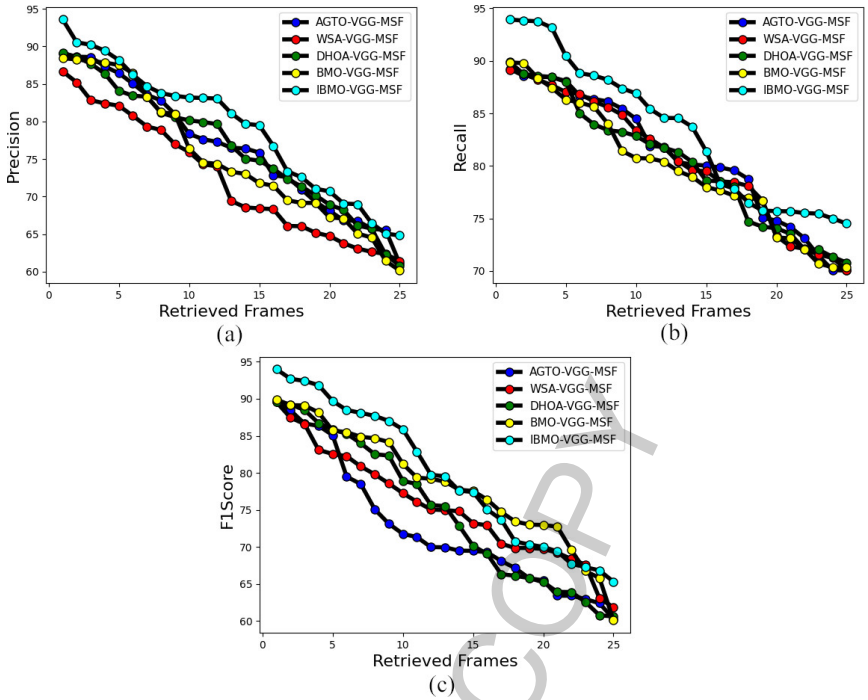

Fig. 10. Comparing the efficiency of the developed video retrieval system on dataset 2 among different heuristic strategies in regards to (a) Precision (b) Recall and (c) F1-score.

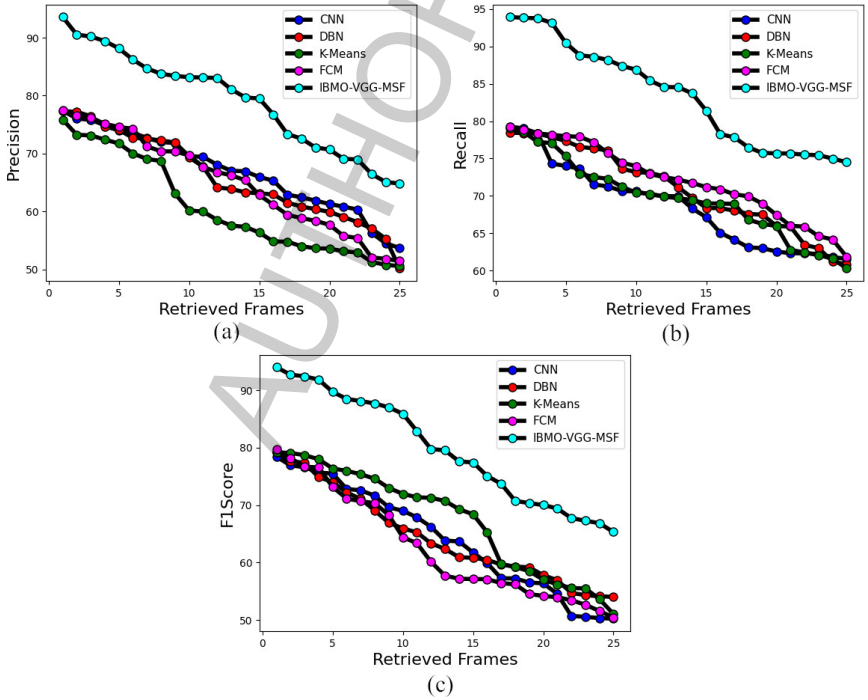

Fig. 11. Comparing the efficiency of the developed video retrieval system on dataset 1 among conventional video retrieval techniques in regards to (a) Precision, (b) Recall, and (c) F1-score.

Table 2
Statistical comparison on developed IBMO-VGG16-MSF-based Video retrieval system using dataset 1 among different heuristic algorithms

Statistical measures	AGTO-VGG -MSF [31]	WSA-VGG-MSF [32]	DHOA-VGG-MSF [33]	BMO-VGG-MSF [27]	IBMO-VGG-MSF
Precision					
Mean	77.32399	74.2583	73.08441	75.98395	**77.41436**
Worst	62.6275	60.4382	60.96886	60.58182	**62.09634**
Median	81.11305	73.54204	73.60272	73.28085	**87.55306**
Standard deviation	7.811105	6.41224	8.067151	8.459064	**8.991385**
Best	89.82156	86.97967	89.02253	89.8396	**89.9425**
Recall					
Mean	79.97084	77.58702	78.01726	79.92997	**80.30999**
Worst	72.24259	71.06397	70.14632	70.59738	**71.00212**
Median	77.88446	79.10719	77.3768 6	79.32937	**79.76744**
Standard deviation	5.400011	5.205283	6.245457	5.078836	**5.687808**
Best	89.63736	89.1667	89.14829	89.4591	**89.89587**
F1-Measure					
Mean	79.97084	80.18702	78.01726	79.92997	**80.30999**
Worst	72.24259	71.06397	70.14632	70.59738	**71.00212**
Median	77.88446	78.10719	77.37686	78.72937	**79.76744**
Standard deviation	5.400011	5.205283	5.245457	5.078836	**5.687808**
Best	89.63736	89.1667	89.14829	89.4591	**89.89587**

Table 3
Statistical comparison of the developed IBMO-VGG16-MSF-based Video retrieval system using dataset 1 among different Baseline Video Retrieval Techniques

Statistical measures	CNN [20]	DBN [21]	K-Means [30]	FCM [29]	IBMO-VGG-MSF
Precision					
Mean	62.81289	67.63174	66.40529	67.62055	**77.41436**
Worst	50.49239	50.99712	50.10335	53.86785	**62.09634**
Median	62.33608	68.96347	67.67879	67.51506	**77.55306**
Standard deviation	9.056142	8.038948	7.959281	7.12709	**8.991385**
Best	79.50884	79.0047	78.90056	79.85306	**89.9425**
Recall					
Mean	69.76423	68.81086	68.12913	71.17877	**80.30999**
Worst	60.43784	60.96844	60.32505	60.4297	**71.00212**
Median	71.42504	68.88246	67.85985	73.66789	**79.76744**
Standard deviation	6.332132	5.887191	5.41279	6.776903	**7.687808**
Best	79.47497	78.73546	79.48488	79.51804	**89.89587**
F1-measure					
Mean	66.10702	64.95361	63.9782	63.641	**76.28259**
Worst	50.68533	50.80106	50.07303	51.66069	**60.30673**
Median	68.86972	64.9355	61.59801	63.19753	**77.82247**
Standard deviation	7.386126	7.215337	7.111548	7.49414	**7.767947**
Best	79.0121	78.94744	78.34077	79.0447	**89.52941**

6.5. Performance analysis using data set 2

The performance analysis of the developed IBMO-VGG16-MFO-based video browsing and video retrieval system adopted with various heuristic algorithms is illustrated in below Fig. 10, and a comparison among various baseline video retrieval systems is depicted in Fig. 11 while considering dataset 2. This performance analysis is carried out following the number of received frames. The newly proposed video retrieval system is 14.70% superior to CNN, 18.18% superior to DBN, 27.86% superior to K-means, and 39.28% superior to FCM, while considering the received frame number is 20 on precision analysis.

Table 4
Statistical comparison of the developed IBMO-VGG16-MSF-based Video retrieval system using dataset 2 among different heuristic algorithms

Statistical measures	AGTO-VGG -MSF [31]	WSA-VGG-MSF [32]	DHOA-VGG-MSF [33]	BMO-VGG-MSF [27]	IBMO-VGG-MSF
Precision					
Mean	72.5581	72.37296	73.4041	74.20252	**75.13364**
Worst	61.07015	61.4056	59.78862	60.14162	**62.81116**
Median	76.48725	69.41532	76.79682	73.29421	**77.08387**
Standard deviation	8.241546	7.983077	8.114271	8.213678	**8.501394**
Best	89.07914	86.64998	89.09859	88.4233	**92.56874**
Recall					
Mean	72.91708	74.55167	78.21558	79.81197	**82.39412**
Worst	70.02577	70.08183	70.78546	70.31729	**73.50329**
Median	78.34048	80.21342	79.32026	79.49927	**80.54648**
Standard deviation	6.105953	6.07143	5.987512	6.045597	**6.780717**
Best	89.22765	88.16222	89.01939	88.8493	**89.93948**
F1-Measure					
Mean	72.51848	74.40866	74.80437	72.42858	**75.65719**
Worst	60.29189	61.85841	60.629	60.10442	**61.3156**
Median	69.95632	74.97271	74.11246	72.82165	**75.53447**
Standard deviation	8.687315	7.221907	9.019824	7.810736	**9.381923**
Best	88.78999	88.55393	87.60373	88.90216	**89.98952**

Table 5
Statistical comparison of the developed IBMO-VGG16-MSF-based Video retrieval system using dataset 2 among different Baseline Video retrieval systems

Statistical measures	CNN [20]	DBN [21]	K-Means [30]	FCM [29]	IBMO-VGG-MSF
Precision					
Mean	66.97647	65.83409	60.65919	65.06784	**75.13364**
Worst	53.68724	50.1468	50.55615	51.44814	**60.81116**
Median	67.03216	63.86588	57.55708	66.22907	**77.08387**
Standard deviation	6.773046	7.436862	8.234399	8.281474	**8.701394**
Best	77.39502	77.32128	75.83875	77.51224	**89.56874**
Recall					
Mean	68.64767	70.91327	69.63499	72.0889	**79.38612**
Worst	61.55512	60.80636	60.27589	61.82065	**70.50329**
Median	69.94785	71.1463	69.68674	72.18091	**80.54648**
Standard deviation	5.598505	5.568761	5.239272	5.035065	**6.780717**
Best	79.26827	78.46547	79.0584	79.29479	**89.93948**
F1-Measure					
Mean	64.22469	64.45412	67.39104	62.60143	**75.65719**
Worst	50.21602	54.01758	51.03091	50.41618	**61.3156**
Median	63.81067	62.3581	70.72978	57.71729	**75.53447**
Standard deviation	9.243475	7.819676	9.08461	9.24318	**9.381923**
Best	78.40012	79.13629	79.1707	79.73401	**89.98952**

Furthermore, the effectiveness of the developed video retrieval system is higher than the other conventional video retrieval systems and other optimization algorithms.

6.6. Statistical analysis using data set 1

Statistical analysis is used to provide better decision support in terms of performance over the developed IBMO-VGG16-MFO-based video browsing and video retrieval system based on median, best, standard deviation, mean, and worst. Comparison among different heuristic algorithms using dataset 1 is given in

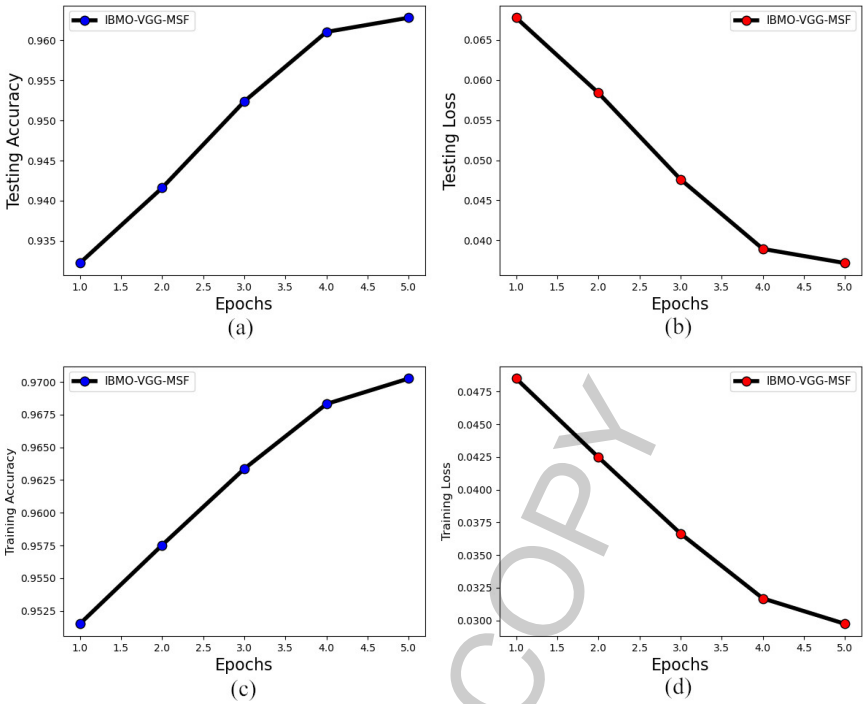

Fig. 12. Performance analysis of the offered model of dataset 1 regarding epochs (a) Testing accuracy, (b) Testing loss, (c) Training accuracy, and (d) Training loss.

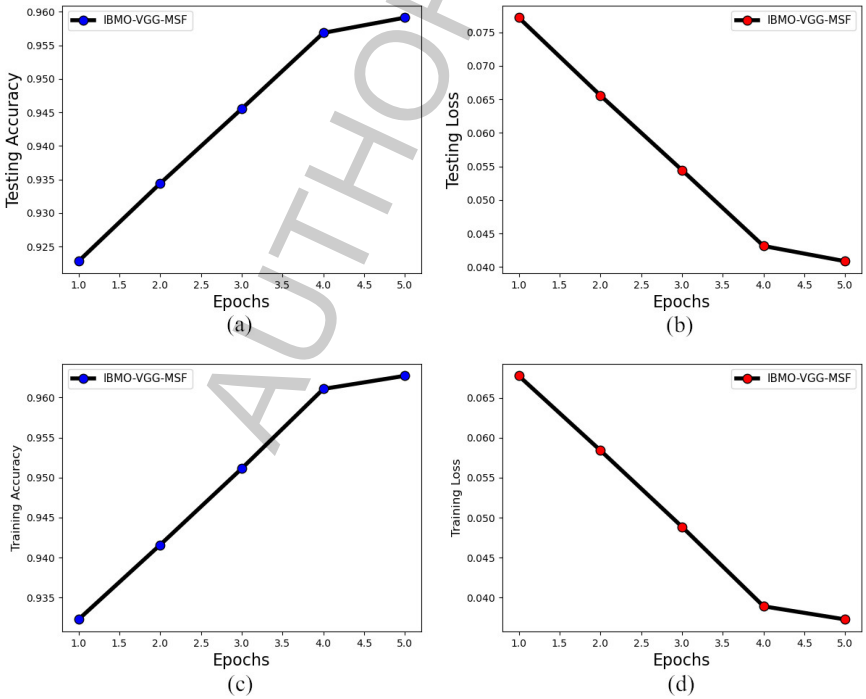

Fig. 13. Performance analysis of the offered model for dataset 2 concerning epochs (a) Testing accuracy, (b) Testing loss, (c) Training accuracy, and (d) Training loss.

Table 2, and among various conventional video retrieval systems is depicted in Table 3. From this statistical analysis, the developed model achieved a precision of 0.42% than AGTO, 3.50% than WSA, 2.93% than DHOA, and 1.47% than BMO concerning the improved mean value. Moreover, the performance of the developed video retrieval process is high in terms of precision, recall, and F1-score.

6.7. Statistical analysis using data set 2

The statistical analysis of the developed video retrieval system using dataset 2 over different heuristic algorithms and different video retrieval systems are depicted in Tables 4 and 5, respectively. The developed model is 18.37% superior to CNN, 21.13% superior to DBN, 6.79% superior to K-means, and 30.86% superior to FCM while regarding the measure of F1-score based on analysis of median value. The overall effectiveness of the developed model is higher than the other heuristic algorithms and baseline video retrieval systems.

6.8. Validation of hyperparameters used in the designed method for dataset 1

The overall performance analysis of the designed IBMO-VGG16-MSF-based Video retrieval system using training loss, testing loss, training accuracy, and testing accuracy for dataset 1 regarding epochs is shown in Fig. 12.

6.9. Validation of hyperparameters in the designed method for dataset 2 regarding epochs

The overall performance analysis of the designed IBMO-VGG16-MSF-based Video retrieval system using training loss, testing loss, training accuracy, and testing accuracy for dataset 2 regarding epochs is shown in Fig. 13.

6.10. Discussion

A few drawbacks of the existing research method are shown here. In practice, monitoring of the overall activity of the moving object in various cameras is quite challenging. Sometimes, it may fail due to the variation of speed and directions of the adjacent moving object. In some cases, it is not effectively focused on object detection and tracking, as well as the background. While screening the entire video, it is not well performed for the complete trajectories. Hence, it contains poor performance with the presence of visual complexities in the video. Thus, a new model is developed for effectively browsing and retrieving the system for video in the wider range of large surveillance systems. Moreover, the developed model can be performed in less amount of time. Hence, it is utilized in larger datasets to validate the effective performance of the designed method. Additionally, it can effectively speed up the training time in the complex network. Moreover, the developed model helps to solve issues like gradient vanishing, overfitting and underfitting, which helps to improve the efficiency of the system performance. While validating with the standard performance metrics, the developed model achieves better performance when compared with other conventional approaches. Here, the PSNR value is maximized with the help of the designed method.

7. Conclusion

A new video condensation method for video browsing and video retrieval system was developed to provide appropriate videos based on user requirements from the large surveillance videos with less utilization of time. The required dataset has been gathered and given to the video condensation stage, where

the background frames were initially extracted, and the YLOv3 network was used to detect objects. Here, the developed IBMO algorithms were used to minimize the object uncovered rate by selecting appropriate frames for frame stitching. The stitched frames were constructed based on the object's activity and the time interval. Finally, the stitched frames were stored in a single video that was the condensed video. In the video browsing and retrieval stage, the condensed video was given as input, and this video was segmented into many portions, followed by deep feature extraction from video segments using VGG16, and these segments were clustered with FCM for segmenting the videos. These clustered segments were stored in the feature database, and this was performed in the training phase. In the testing phase, the user was given queries to get specific contents that were stored in the database, and the features were extracted from the queries and fed to MSF over previously stored clustered segments concerning Euclidean distance and cosine similarity. Finally, the appropriate content was provided for users based on their queries very quickly. The experimental results were compared with previously used video retrieval systems, and the mean value of the developed model is 17.80% superior to CNN, 17.38% superior to DBN, 12.26% superior to K-means, and 20.85% superior to FCM according to the F1-score. The efficacy of the developed video retrieval system was excessively high when compared to other algorithms and conventional video retrieval systems.

References

[1] Xu P, Liu K, Xiang T, Hospedales TM, Ma Z, Guo J, Song Y-Z. Fine-Grained Instance-Level Sketch-Based Video Retrieval. IEEE Trans Circuits Syst Video Technol. 2021; 31(5): 1995-2007.

[2] Yoon H, Han J-H. Content-Based Video Retrieval With Prototypes of Deep Features. IEEE Access. 2022; 10: 30730-30742.

[3] Wang Y, Nie X, Shi Y, Zhou X, Yin Y. Attention-Based Video Hashing for Large-Scale Video Retrieval. IEEE Trans Cognit Dev Syst. 2021; 13(3): 491-502.

[4] Choi YR, Kil RM. Face Video Retrieval Based on the Deep CNN With RBF Loss. IEEE Trans Image Process. 2021; 30: 1015-1029.

[5] Qi M, Qin J, Yang Y, Wang Y, Luo J. Semantics-Aware Spatial-Temporal Binaries for Cross-Modal Video Retrieval. IEEE Trans Image Process. 2021; 30: 2989-3004.

[6] Shi Y, Wei Z, Ling H, Wang Z, Shen J, Li P. Person Retrieval in Surveillance Videos Via Deep Attribute Mining and Reasoning. IEEE Trans Multimedia. 2021; 23: 4376-4387.

[7] Karpenko A, Aarabi P. Tiny Videos: A Large Data Set for Nonparametric Video Retrieval and Frame Classification. IEEE Trans Pattern Anal Mach Intell. 2011; 33(3): 618-630.

[8] Erol B, Kossentini F. Shape-based retrieval of video objects. IEEE Trans Multimedia. 2005; 7(1): 179-182.

[9] Cotsaces C, Nikolaidis N, Pitas I. Face-Based Digital Signatures for Video Retrieval. IEEE Trans Circuits Syst Video Technol. 2008; 18(4): 549-553.

[10] Hoi SCH, Lyu MR. A Multimodal and Multilevel Ranking Scheme for Large-Scale Video Retrieval. IEEE Trans Multimedia. 2008; 10(4): 607-619.

[11] Hu W, Xie D, Fu Z, Zeng W, Maybank S. Semantic-Based Surveillance Video Retrieval. IEEE Trans Image Process. 2007; 16(4): 1168-1181.

[12] Kang EK, Jahng SG, Choi JS. A new indexing method for video retrieval using the rosette pattern. IEEE Trans Consum Electron. 2000; 46(3): 780-784.

[13] Araujo A, Girod B. Large-Scale Video Retrieval Using Image Queries. IEEE Trans Circuits Syst Video Technol. 2018; 28(6): 1406-1420.

[14] Bruno E, Moenne-Loccoz N, Marchand-Maillet S. Design of Multimodal Dissimilarity Spaces for Retrieval of Video Documents. IEEE Trans Pattern Anal Mach Intell. 2008; 30(9): 1520-1533.

[15] Castañón G, Elgharib M, Saligrama V, Jodoin P-M. Retrieval in Long-Surveillance Videos Using User-Described Motion and Object Attributes. IEEE Trans Circuits Syst Video Technol. 2016; 26(12): 2313-2327.

[16] Yang H, Meinel C. Content Based Lecture Video Retrieval Using Speech and Video Text Information. IEEE Trans Learn Technol. 2014; 7(2): 142-154.

[17] Ho Y-H, Lin C-W, Chen J-F, Liao H-YM. Fast coarse-to-fine video retrieval using shot-level spatio-temporal statistics. IEEE Trans Circuits Syst Video Technol. 2006; 16(5): 642-648.

[18] Chiang C-C, Yang H-F. Quick browsing and retrieval for surveillance videos. Multimed Tools Appl. 2015; 74: 2861-2877.

[19] Ding S, Li G, Li Y, Li X, Zhai Q, Champion AC, Zhu J, Xuan D, Zheng YF. SurvSurf: human retrieval on large surveillance video data. Multimed Tools Appl. 2017; 76: 6521-654.

[20] Lin F-C, Ngo H-H, Dow C-R. A cloud-based face video retrieval system with deep learning. J Supercomput. 2020; 76: 8473-8493.
[21] Poornima N, Saleena B. An automated approach to retrieve lecture videos using context-based semantic features and deep learning. Sādhanā. 2020; 45(254): na-na.
[22] Kumar BS, Seetharaman K. Content based video retrieval using deep learning feature extraction by modified VGG_16. J Ambient Intell Human Comput. 2022; 13: 4235-4247.
[23] Ullah A, Muhammad K, Hussain T, Baik SW, De Albuquerque VHC. Event-Oriented 3D Convolutional Features Selection and Hash Codes Generation Using PCA for Video Retrieval. IEEE Access. 2020; 8: 196529-196540.
[24] Nguyen HT, Jung S-W, Won CS. Order-Preserving Condensation of Moving Objects in Surveillance Videos. IEEE Trans Intell Transp Syst. 2016; 17(9): 2408-2418.
[25] Zhu J, Feng S, Yi D, Liao S, Lei Z, Li SZ. High-Performance Video Condensation System. IEEE Trans Circuits Syst Video Technol. 2015; 25(7): 1113-1124.
[26] Oguine KJ, Oguine OC, Bisallah H. YOLO v3: Visual and Real-Time Object Detection Model for Smart Surveillance Systems (3s). arXiv preprint arxiv:2209.12447. 2022.
[27] Mahmood M, Al-Khateeb B. The Blue Monkey: A New Nature Inspired Metaheuristic Optimization Algorithm. Period Eng Nat Sci (PEN). 2019; 7(3): 1054-1066.
[28] Desai P, Pujari J, Sujatha C, Kamble A, Kambli A. Hybrid Approach for Content-Based Image Retrieval using VGG16 Layered Architecture and SVM: An Application of Deep Learning. SN Comput Sci. 2021; 2: 170.
[29] Xu H, Yao S, Li Q, Ye Z. An Improved K-means Clustering Algorithm. In: 2020 IEEE 5th International Symposium on Smart and Wireless Systems within the Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS). 2020.
[30] Pei H-X, Zheng Z-R, Wang C, Li C-N, Shao Y-H. D-FCM: Density based fuzzy c-means clustering algorithm with application in medical image segmentation. Procedia Comput Sci. 2017; 122: 407-414.
[31] Abdollahzadeh B, Gharehchopogh FS, Mirjalili S. Artificial gorilla troops optimizer: A new nature-inspired metaheuristic algorithm for global optimization problems. Int J Intell Syst. 2021; 1: 72.
[32] Kaveh A. Water Strider Optimization Algorithm and Its Enhancement. Advances in Metaheuristic Algorithms for Optimal Design of Structures. 2021; 783-848.
[33] Kanna SKR, Sivakumar K, Lingaraj N. Development of Deer Hunting linked Earthworm Optimization Algorithm for solving large scale Traveling Salesman Problem. Knowledge-Based Syst. 2021; 227: 107199.