Segmentation And Feature Extraction Of Content Based Video Retrieval Using Maximum Likelihood Regression Model With Modified-Vgg-16

[1*]B. Satheesh Kumar , [2]K. Seetharaman

[1*]Research Scholar, Department of computer science and engineering,, Annamalai University, Annamalai Nagar.

[2]Professor, Department of computer and information science, Annamalai University, Annamalai Nagar.

Corresponding Email id : vbsatheeshkumarphd@gmail.com

Received Date: June 06, 2021; **Revised Date**: June 06, 2021; **Accepted Date:** November 12, 2021
Publication Date: December 12, 2021 **-ISSN**: 1735-188X

Abstract

The recent challenge faced by the users from the multimedia area is to collect the relevant object or unique image from the collection of huge data. During the classification of semantics, the media was allowed to access the text by merging the media with the text or content before the emergence of content based retrieval. The identification of this feature has become major challenges, so to overcome this issue this paper focuses on a deep learning technique with maximum likelihood regression (MLR) model for segmentation and Feature extraction of the input video. Likelihood estimation is to roughly measure the level of pixel, and then regression method determines pixel level to certainly transformblurred and unwanted pixels. The segmentation is done based on the likelihood estimation and the feature of this segmented video is extracted using Modified_ VGG-16 (M_VGG-16) architecture. The result of this technique has been compared with the existing other feature extraction techniques such as conventional Histogram of Oriented Gradients (HOG), LBP (Local Binary Patterns) and CNN (Convolution Neural Network) methods. In this scheme the video frame image retrieval is performed by assigning the indexing to the all video files in the database in order to perform the system more efficiently. Thus the system produces the top result matches for the similar query in comparison with the existing techniques based on precision of 90%, recall of 93% and F1 score of 91% in optimized video frame retrieval.

Keywords: content based retrieval, segmentation, Feature extraction, MLR, M_VGG-16

1. INTRODUCTION

In modern era, storage environments of computer allows the user to store and retrieve the digital data with huge size which also make the complexity during the retrieval relevant data or information based on demand. So it is highly difficult for the users to retrieve the data when proper tools or techniques are not available [1]. Furthermore when it comes to the video files, the database allows storing large size of video with irrespective of size and type in its database and it makes the users to face challenging issues during the retrieval of relevant video.

The video retrieval systems (VRS) with Video Content Retrieval (VCR) functionality do not satisfy the users. To satisfy the users instead of raw video data they want to query the content. For example, if a user wants to analyze the specific events instead on complete event such as soccer (retrieving the goal event instead of retrieving the whole game) then content based search retrieval becomes the major challenging problem. Therefore, numeric and textual data are managed by the tools which manipulate video content similar to that of traditional databases [2].

Several VRS are available related to the spatial features like shape, texture and color. But video related systems are different from image and the literature is scarce. Regarding video indexing and retrieving in compressed domain, only limited research exists and moreover the operation is performed only at the frame level without taking into account the video content [3]. Features like motion vectors as well as DCT (Discrete Cosine Transform) coefficients are extracted from the compressed Moving Picture Experts Group (MPEG) video and are utilized. Hence the issues of high cost during decompression and operating at pixel level are overcome. The video content is provided by the characteristics of the object in the video. Hence for retrieving and indexing a video efficiently, there arises a necessity to describe the video frame images. Only few works are done in pixel domain and hence the computational cost is higher [4].

For retrieving video frames, the existing methods use similar video frames where Euclidean distance is computed among the features of frame which is a complex factor as numerous frames exists; thereby complexity is increased and retrieval efficiency is limited. Convolutional Neural Network based Hashing (CNNH) was introduced to improve the retrieving efficiency and overcome complexity that occurs due to Euclidean distance. The CNNH integrates hashing technique and DNNs to increase the efficiency of the model. The focus of video recovery is to identify a particular object from a video frame given as input based on few identified frames of the video database. In general, retrieving a video frame is the process of identifying identical video frames based on query which is sent back to the user. Video retrieval system replaces the low-level features with high-level features thereby the ability of matching characters in video frames was improved. The high-level features are obtained efficiently using Deep Neural Network (DNNs) [5].

Likelihood regression is proposed in this paper to estimate the overall pixel intensity of the input video sequence [6]. After estimation, its relative pixels are also estimated. By this approach, specific pixel values are converted into a video sequence. The estimated video frames are then stored in the database for further processing.M_VGG-16 architecture is used to extract the features from the pre-processed video [7]. The results of this model are comparatively evaluated with existing methods and proved that the proposed model is efficient than the methods considered for comparison.

Paper organization is as follows: The existing work related to video retrieval is discussed in section 2. The proposed architecture with its working is elaborated in section 3. Performance metrics along with experimental results and comparison is illustrated in section 4 while conclusions of this work in section 5.

2. RELATED WORKS

Few existing content based video retrieval (CBVR) systems are discussed in this section. In the past century, several projects were developed by institutes which focused on accessing digital video intelligently [8]. The project of Carnegie Mellon University named as Infor media Digital Video Library and Columbia University‘s Video Q project was the first on-line video search engine which supported automated indexing based on object and spatio-temporal queries [9]. The focus was on processing videos, particularly, detection of shot boundary, summarization and segmentation of video, detecting objects from video and CBVR [10]. Various recent approaches involved in CBVR are presented in the literature review [11]. In [12], VRS was developed based on motion histogram and Apache Hadoop. In [13], a distributed real-time processing model was developed. These above mentioned works made use of entire video for creating and comparing video signatures and thus the processing time was reduced.

Fabletet.al. introduced CBVR and indexing with the objective to provide overall explanation of dynamic content of video shots with no motion segmentation and also with no usage of thick optic flow fields. Spatio-temporal distribution was developed. Hierarchical structure for the video database processed was built in accordance with the similarity of motion content. As a result, binary tree was obtained where every node was connected with estimated causal Gibbs model. Subsequently, query-by-example retrieval was carried out with this tree by using maximum a posteriori (MAP) condition [14].

Fermanet.al. developed histogram-based color descriptors for capturing and representing color properties of several images. Alpha-trimmed average histograms provided single frame or image histograms jointly into a filter for generating robust and reliable color histograms. By this, color variations, effects of brightness, occlusion as well as editing color effects were eliminated. This approach outperformed key frame-based techniques. Intersection histogram was designed to reflect the number of pixels for a specified color. It was proved that this method was fast and efficient in identifying video segment of the query frame[15].

Sahooet.al. coined an approach to quickly detect slow moving objects from a video which integrated spatial along with temporal information. A simple and easier temporal segmentation method termed as average frame differencing was developed to minimize error during classification. Moreover, time complexity was also reduced as valley-based spatial segmentation was involved. This method when tested with different videos taken from the available public database proved its efficiency in CPU time and outperformed JSEG, Otsu, Zhu and Singla's method by reducing misclassification error. This error was high for sudden change in the scene or light variation in background or television. It was suggested that integrating code book-based background modeling along with temporal differencing approach would provided better efficiency for video information [16].

Roy et. al. designed date spotting based information retrieval system. This was used to deal with natural scene images as well as video frames which contained texts as complex backgrounds. Moreover, line based date spotting method with HMM (Hidden Markov Model) was employed for detecting date information present in the given text. If it was in RGB, gray image conversion was performed efficiently using wavelet decomposition and gradient sub-bands for enhancing information. From the gray and binary images, PHOG (Pyramid HOG) feature was extracted [17].

Muhlinget. al. coined a deep learning method for supporting professional media production. Specifically, algorithms like similarity search, detection, clustering and recognition of face were integrated in a multimedia for inspecting as well as retrieving video information effectively. Concept detection is incorporated with similarity search in a multi-task learning method in order for sharing network weights thereby approximately half of the time taken for computation was saved. Further, visual concept lexicon was introduced for quick fast video retrieval. This approach proved its efficiency [18].

The standard techniques require a reference frame, previous or optical flow mask to perform segmentation and feature extraction. The increasing changes in object appearance prevent the model to produce accurate results of segmentation.

3. SYSTEM MODEL

Overall Architecture of Proposed Methodology

The overall system architecture of the proposed methodology has been given in figure-1.

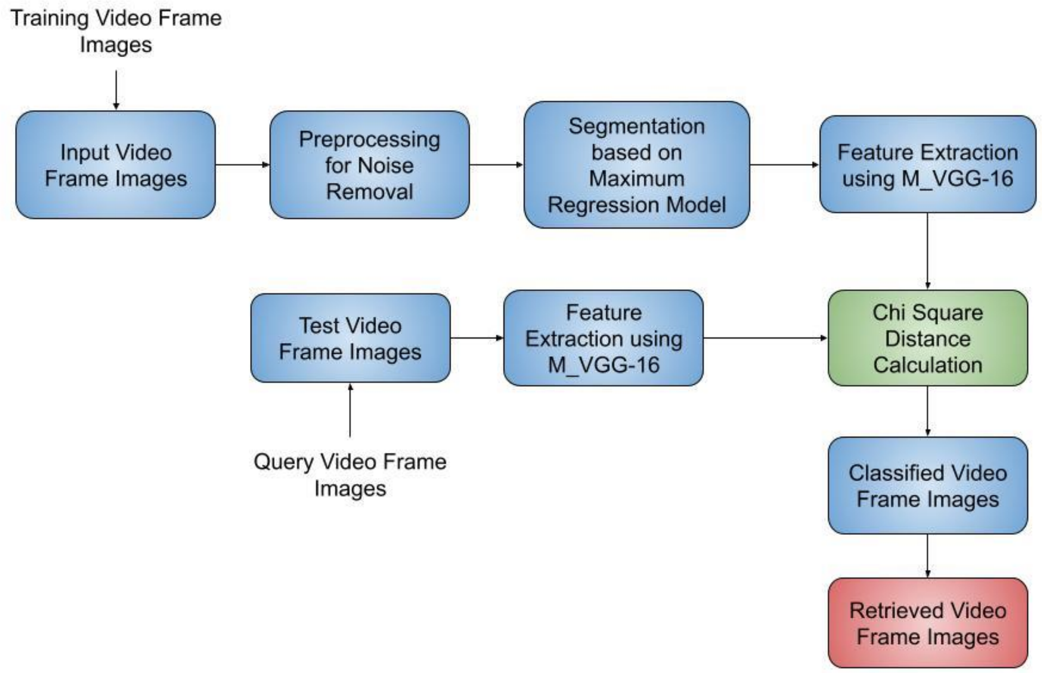

Training Video Frame Images
Input Video Frame Images
Preprocessing for Noise Removal
Segmentation based on Maximum Regression Model
Feature Extraction using M_VGG-16
Test Video Frame Images
Query Video Frame Images
Feature Extraction using M_VGG-16
Chi Square Distance Calculation
Classified Video Frame Images
Retrieved Video Frame Images

Figure 1: Overall System Architecture of MLR_MVGG-16

Initially the input video frame images have been taken and it is pre-processed for noise removal. After the noise removal, video frames have been segmented using maximum likelihood regression model (MLR). Then the features of image have been extracted using MVGG_16. Then these extracted features have been classified. Then in query set of the image, the test image has been taken and again its features have been extracted using MVGG_16. Then its chi squared distance has been calculated which then obtain classified test data of the input image. Finally the similarity measures of image from training set and query set has been compared and there it retrieves the enhanced video frames.

Segmentation based on maximum likelihood regression model (MLR):

This likelihood estimation method estimates the pixel range effectively from video sequence frame thereby retrieving video frame. In contrast to the estimated probability, converted frame exhibits minimal probability. Video sequences are estimated from video frame to process further using a regression model where pixel conversion takes place exhibiting minimal density for every pixel. At last using this proposed model, video frame is extracted and retrieved

Step 1: Video sequence is taken as input

Step 2: Likelihood estimation model is designed

Step 3: Regression model is constructed

Step 4: Segmenting video frame and extraction of features

Step 5: Segmentation of frame sequences

As per the above processing steps, likelihood regression estimation model is employed in evaluation. The processing steps of likelihood regression estimation are summarized below.

Likelihood regression estimation model takes input video frame as input and video frame is retrieved as output.

```
1. Video frame extraction about definite time
2. Likelihood ratio is estimated for video sequence
3. Video sequence is converted to frame
4. For the converted frame, pixel intensity is estimated
5. for Iran ging from 0 to 255
            I=I+ pixel intensity
      else I = 0
End for;
6. Video frame sequence given as input is applied
7. A dataset is created with video sequence given as input
8. Query frame is sent
9. Frame is segmented
```

With this model, pixel intensity is determined and processed which are measured by considering pixel intensity approximation [19].

From extracted video frame, pixels are retrieve using maximum likelihood estimation and then regression model is applied where minimal pixel range are estimated. With no limits in RM generalization, digital video frame of every scene analyzed is assumed as matrices with predetermined size, $A \equiv A_{m,n}$

Rectangular video frame P is considered representing first quadrant of plane x and y as n-1 and m-1 accordingly. For a video frame, geometric feature is analytically defined over P which is given as z = z(x, y) where $z(i, j) = a_{ij}$

A sample$a_{ij}$is assumed which includes random variable $X_{ij} = z(i, j) + X$ here X denotes random mean value EX = 0 and variance$DX = \sigma^2$. Let$z(i, j) = a_{ij}$ as σ^2 is small. In the video frame, to analyse the variations in pixels these parameters are evaluated. When the selected pixel frame is closer to the basic function, maximum likelihood$\{f_0(x, y),\ f_1(x, y), \dots \dots, f_k(x, y)\}$is used where elements in P with unknown values of z(x, y), in a independent system with(x, y) ∈ P and linear combination of$f_k(x, y)$function using suitable unknown coefficientθ'_k is represented by eq. (1):

$\sum_{k=0}^{K} \theta'_k f_k(x, y)$ (1)

Here, θ'_k gives approximate z(x, y) and least square element method is used for estimation of $a_{ij}$for matrix $A \equiv A_{m,n}$ and θ_k^{Λ} values are estimated. The matrix is defined by eq. (2),

$F_m n, K + 1 = (f_{sk})$ (2)

Where, $ggg(f_{s0}, f_{s1}, \dots \dots, f_{sK}) = (f_0(i, j), f_1(i, j), \dots, f_K(i, j))$and $s = in + j$. Using the below equation, non-singular matrix is derived by eq. (3).

$M = F^T F$ (3)

(K + 1) ∗ (K + 1) is the dimension and the transpose operation is denoted by T. $\overset{\Lambda}{\theta}$of the vector θ' is estimated with the following equation (4).

$\overset{\Lambda}{\theta} = M^{-1} F^T a$ (4)

In the video sequence, digital frame includes larger matrix dimensions and thus are divided as sub-matrices whose size is small$A \equiv A_{m,n}$. By this, the frame size of the feature set for every pixel frame has to be smaller than the size of sub-matrix. The m and n values has to be small and in the first state of regression method, it is implemented to estimate regression coefficient.

As the video sequences S1, S2……Sa are used with K1, K2, …, Knclasses, then the estimated vectors are T1, T2, …, Tn. Here, the cardinality sets Si and Tiare similar and are preferred for huge sets. When cardinality for the video frame is matched then the proposed model processes the video frame and video frames are retrieved from large datasets.

Modified VGG_16

MVGG_16 is prediction models that can achieve 90% top-5 test accuracy. The overall structure of MVGG_16 is shown in Figure-2. cov1 layer takes RGB image as input whose fixed size is 224 x 224 and passes through several conv layers and the filters with a small receptive field 3×3 are used. Even 1×1 conv filters is used in one configuration which is assumed to be input channels with linear transformation and the stride is set as 1. After convolution, spatial padding of conv. layer input preserves spatial resolution. Five max-pooling layers perform spatial pooling over 2×2 pixel window, whose stride is fixed as 2.

Convolution layers are succeeded by Three Fully-Connected (FC) layers with two layers with 4096 channels each and the final one containing 1000 channels carries out 1000-way ILSVRC classification. The configuration of FC layers is identical for every network. Finally, the model has soft-max layer.

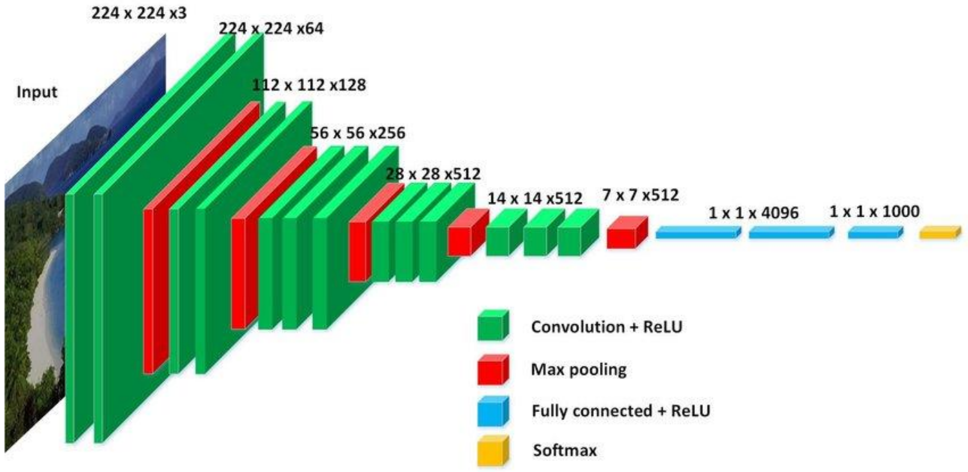


Figure.2: Overview of MVGG_16 Structure

Every hidden layer has ReLU non-linearity and noticed that networks do not contain Local Response Normalisation (LRN). Moreover, it is noted that this provides no improvement in the performance of ILSVRC dataset and consumes more memory and takes much time for computation.

Table1outlines Conv Net configurations. The names A-E refer the nets. Every configuration is based on the generic design but varies with depth: from 11 to 19 weight layers in the network A (8 conv. and 3 FC layers) to E (16 conv. and 3 FC layers). For the conv. Layers, the number of channels (width) ranges from 64 to 512 which increases by 2 after every max-pooling layer.

Table-1.ConvNet Configuration

Conv Net Configuration					
A	A-LRN	B	C	D	E
11 Weight Layers	11 Weight Layers	13 Weight Layers	16 Weight Layers	16 Weight Layers	19 Weight Layers
Input (224 x 224 RGB image)					
Conv3-64	Conv3-64	Conv3-64	Conv3-64	Conv3-64	Conv3-64
	LRN	Conv3-64	Conv3-64	Conv3-64	Conv3-64
Max pool					
Conv3-128	Conv3-128	Conv3-128	Conv3-128	Conv3-128	Conv3-128
		Conv3-128	Conv3-128	Conv3-128	Conv3-128
Max pool					
Conv3-256	Conv3-256	Conv3-256	Conv3-256	Conv3-256	Conv3-256
Conv3-256	Conv3-256	Conv3-256	Conv3-256	Conv3-256	Conv3-256
			Conv1-256	Conv3-256	Conv3-256
					Conv3-256
Max pool					
Conv3-512	Conv3-512	Conv3-512	Conv3-512	Conv3-512	Conv3-512

Conv3-512	Conv3-512	Conv3-512	Conv3-512	Conv3-512	Conv3-512
			Conv1-512	Conv3-512	Conv3-512
					Conv3-512
Max pool					
Conv3-512	Conv3-512	Conv3-512	Conv3-512	Conv3-512	Conv3-512
Conv3-512	Conv3-512	Conv3-512	Conv3-512	Conv3-512	Conv3-512
			Conv1-512	Conv3-512	Conv3-512
					Conv3-512
Max pool					
FC-4096					
FC-4096					
FC-1000					
Soft max					

ImageNet dataset comprises of images whose size is fixed to 224*224 having RGB channels. A tensor of (224, 224, 3) is taken as input. After the input image is processed by the proposed model, a vector with 1000 values is produced as output by eq. (5).

$$\hat{y} = \begin{bmatrix} \hat{\mathcal{y}}_0 \\ \widehat{y_1} \\ \widehat{y_2} \\ \widehat{y_3} \\ \cdot \\ \cdot \\ \cdot \\ , \\ .\, y_{999} \end{bmatrix} \qquad (5)$$

For the corresponding class, classification probability is represented by this vector. For the model which predicts that the image with probability 0.1, 0.05,0.05, 0.03belongs to class 0, class 1, class 2 and class 3 respectively and class 780 and class 999 with probability 0.72, and 0.05respectively while other class have a probability 0. Then classification vector is by eq. (6):

$$\hat{y} = \begin{bmatrix} \hat{y}_0 = 0.1 \\ 0.05 \\ 0.05 \\ 0.03 \\ \cdot \\ \cdot \\ \cdot \\ y_{780} = 0.72 \\ \cdot \\ \cdot \\ y_{999} = 0.05 \end{bmatrix} \qquad (6)$$

Softmax function is used to ensure that these probabilities add to 1 which his defined by eq. (7):

$$P\left(y = j \mid \Theta^{(i)}\right) = \frac{e^{\Theta^{(i)}}}{\sum_{j=0}^{k} e^{\Theta_k^{(i)}}} \qquad (7)$$

The 5 most probable candidates are taken into vector by eq. (8).

$$C = \begin{bmatrix} 780 \\ 0 \\ 1 \\ 2 \\ 999 \end{bmatrix} \qquad (8)$$

and the ground truth vector is given by eq. (9):

$$G = \begin{bmatrix} G_0 \\ G_1 \\ G_2 \end{bmatrix} = \begin{bmatrix} 780 \\ 2 \\ 999 \end{bmatrix} \qquad (9)$$

The Error function is then defined by eq. (10):

$$E = \frac{1}{n} \sum_k \min_i d(c_i, G_k)$$
$$\text{where } d = 0 \text{ if } c_i = G_k \text{ else } d = 1 \qquad (10)$$

Hence, for this example, the loss function is given by eq. (11) and by eq. (12):

$$E = \frac{1}{3}\left(\min_i d(c_i, G_1) + \min_i d(c_i, G_2) + \min_i d(c_i, G_3)\right) \qquad (11)$$

So,

$$E = \frac{1}{3}(0 + 0 + 0) \qquad (12)$$

E=0, as every category in ground truth exists in the top-5 predicted matrix, loss becomes 0.

Chi-square Distance Calculation

This distance is derived from Pearson's Chi-squared test statistic$\chi^2(x, y) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i}$. This is normally used for comparing two discrete probabilitydistributions. Moreover, the function d(x; y) must be symmetric for objects x and y. considering real vectors $X = (x_1, x_2, \dots \dots, x_n)$ and$Y = (y_1, y_2, \dots \dots, y_n)$, Chi-square distance measure is given using the following formulaby eq. (13):

$$d(x, y) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i} \qquad (13)$$

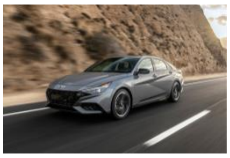
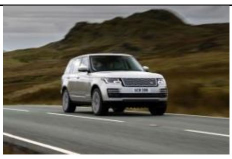
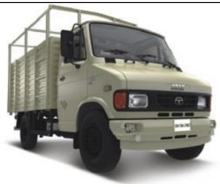
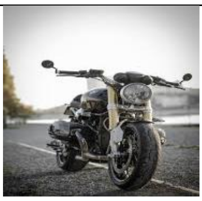
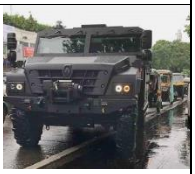
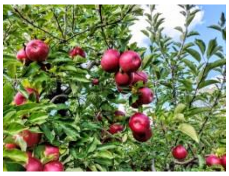
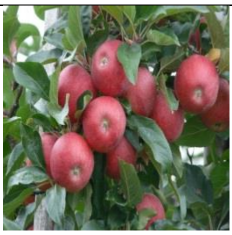
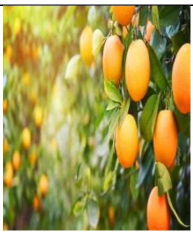
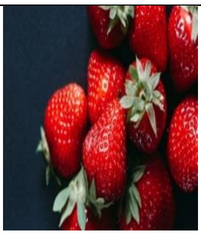
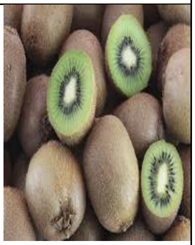
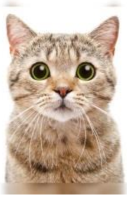
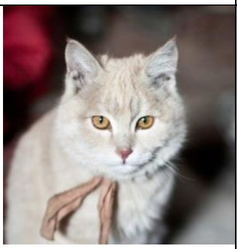
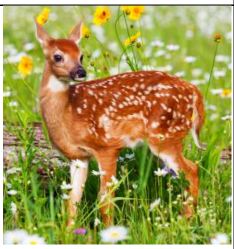
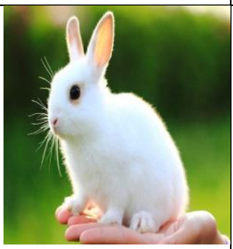
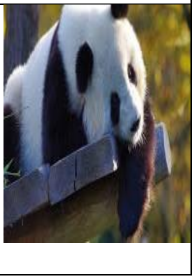
In order to avoid zero in denominator, a revised version is coined by eq. (14):

$$d(x, y) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{2 + x_i + y_i} \qquad (14)$$

4. PERFORMANCE ANALYSIS

The performance analysis for the proposed MLR_MVGG_16 is discussed here. Table 2 shows the video retrieval from the input image using the existing and proposed techniques. MLR_MVGG_16 shows better performance in retrieving video by giving the optimized output with higher efficiency.

Table 2- Frame Retrieval from Input Image using Various Techniques

Query Image	**Retrieved Output Frame**			
	MLR_MVGG_16	**CNN**	**LBP**	**HOG**
Car image 				
Apple image 				
Cat image 				

The metrics considered to evaluate the efficiency of this proposed model are Precision, Recall, and F1 Score which are defined below.

Precision is the ratio of True Positive which is correct Real Positives and is given by eq. (15)

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \qquad (15)$$

Recall is the ratio of the true Positives which are correct Predicted Positive and is given by eq. (16)

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False negative}} \quad (16)$$

F1 score is a metric which denotes accuracy. Both precision and recall are considered to compute the score which is defined as by eq. (17)

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (17)$$

Parametric Analysis of the Proposed and Existing methods

Table 3: Values Obtained for different Parameters and methods

Query Image	Video	No. of Frames	Method	No. of relevant Retrieved Frames	No. of relevant Predicted Frames	No. of relevant video frame images in database	Precision (%)	Recall (%)	F1 Score (%)
Car	0.700k	1801	MLR_MVGG_16	276	268	296	90.54	93.24	91.870
			CNN	93	268	387	69.25	24.03	35.679
			LBP	12	268	502	53.38	0.023	0.045
			HOG	54	268	644	41.61	0.083	0.165
Apple	0.700k	1801	MLR_MVGG_16	302	598	634	94.32	94.95	94.63
			CNN	180	260	440	59.09	40.90	48.34
			LBP	97	184	281	65.48	34.51	45.20
			HOG	68	140	208	67.30	32.69	44.00
Cat	0.700k	1801	MLR_MVGG_16	1690	1456	1801	80.8	93.83	86.85
			CNN	650	722	1601	45.09	53.09	48.76
			LBP	487	612	1510	40.52	32.25	35.91
			HOG	287	378	1887	20.03	15.20	17.29

Table 3 shows the comparison of proposed MLR_MVGG_16 method and existing CNN, LBP and HOG techniques. The number of retrieved frames, predicted frames and relevant video frame images are analyzed and their precision, recall and F1_score are estimated. For example, when the no. of frames is 1801 for query image car, MLR_MVGG_16 obtains 276 and 268 number of relevant retrieved frames and predicted frames respectively; number of relevant video frame images in database is 296, the precision obtained is 90.54%, recall is 93.24%, and F1 Score is 91.87%. The results of CNN, LBP, and HOG are not optimized as proposed MLR_MVGG_16. The results of the proposed architecture are presented as graphs below.

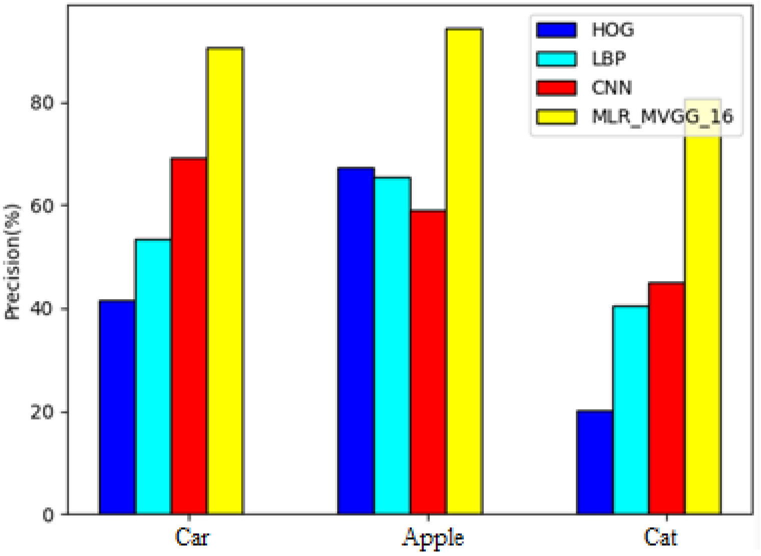


Figure.3: Comparison of Precision

The above figure 3 compares the precision obtained for existing and proposed techniques for various video frame images. This shows that the proposed MLR_MVGG_16 gives better precision percentage i.e. above 80% while all the existing techniques (CNN, LBP, and HOG) give the precision percentage ranging from 20% to 70%.

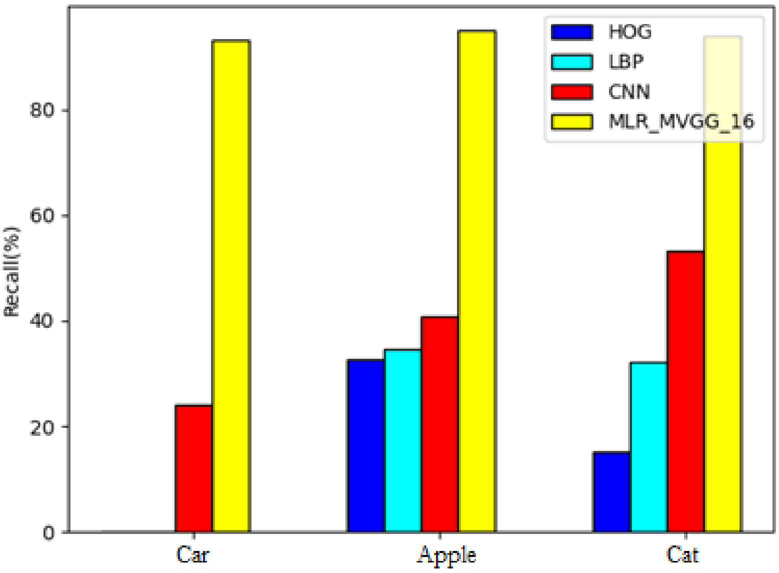


Figure 4: Comparison of Recall

The above figure 4 compares the recall values of existing and proposed techniques for various images. This shows that the proposed MLR_MVGG_16 gives better recall percentage i.e. above 80%. The existing technique CNN gives recall percentage between 20% to 60% and other two existing techniques (LBP, HOG) have no improvement in recall percentage.

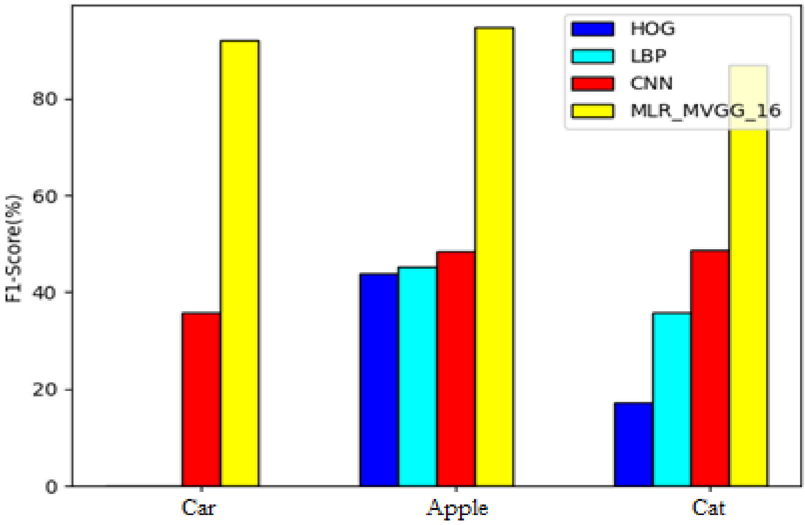


Figure 5: Comparison of F1 Score

The above figure 5 compares F1 score of the existing and proposed techniques for car, apple, and cat. This shows that the proposed MLR_MVGG_16 gives better F1 score percentage i.e. above 80%. The existing technique CNN gives F1 score from 30% to 50%, other two existing techniques (LBP, HOG) has no improvement in F1 Score.

5. CONCLUSION

In spite of the several advanced video retrieval algorithms, a small impact exists on content based video retrieval research for commercial applications with few exceptions like video segmentation. The open issue lies in choosing that features which reflects the real interest of human. So this paper proposed MLR_MVGG_16 deep learning technique for segmentation and feature extraction. A likelihood regression model MLR_MVGG_16 is introduced for frame retrieval from video sequences whose performance is analysed with the existing techniques. From the experiment, the results prove the performance of the proposed model is efficient in terms of precision of 90%, recall of 93% and F1 score of 91%.As a future work, this proposed model can be involved in medical applications, signal processing, satellite communication and signal processing as these fields are challenging in processing video as videos are complex and most of the frames are similar. Moreover, proposed model in integration with classifiers has to be evaluated to estimate the improvement in frame retrieval.

REFERENCES

[1] Mounika, B. R., & Khare, A. (2020, January). Content based video retrieval using histogram of gradients and frame fusion. In Twelfth International Conference on Machine Vision (ICMV 2019) (Vol. 11433, p. 114332J). International Society for Optics and Photonics.

[2] Saritha, R. R., Paul, V., & Kumar, P. G. (2019). Content based image retrieval using deep learning process. Cluster Computing, 22(2), 4187-4200.

[3] Haojin Yang; Christoph Meinel, "Content Based Lecture Video Retrieval Using Speech and Video Text Information", IEEE transactions on learning technologies, Vol. 7 No. 2, 2014, pp. 142- 154.

[4] Dr. H.B. Kekre, Dr. Dhirendra Mishra, Ms. P. R. Rege, “Survey on Recent Technique in Content Based Video Retrieval”, International Journal of Engineering and Technical Research (IJETR), (vol3), pp.69-73,2015.

[5] Aparajeeta J, Mahakud S, Nanda PK, Das N (2018) Variable variance adaptive mean-shift and possibilistic fuzzy C-means based recursive framework for brain MR video frame segmentation. Expert Syst Appl 92: 317–333.

[6] Sathiyaprasad, B., Seetharaman, K., & Kumar, B. S. (2020). Histogram-based Threshold Segmentation of Video Frames using Otsu’s Method. AIJR Abstracts, 25.

[7] B. Sathiyaprasad and K. Seetharaman (2020). Medical Surgical Video Recognition and Retrieval Based on Novel Unified Approximation, J. Med. Imaging Health Inf. Vol. 11, No. xx, 2021

[8] Araujo A, Girod B (2017). Large-scale video retrieval using image queries. IEEE Trans CircSyst Video Technol 28(6):1406–1420

[9] Bouyahi M, Ayed YB (2020). Video scenes segmentation based on multimodal genre prediction. Procedia Comput Sci 176:10–21

[10] B. Sathiyaprasad,K. Seetharaman (2019).Text-Image Queries based Video Retrieval using Image Ontology Queries Formation. Jour of Adv Research in Dynamical & Control Systems, vol. 11, no.8, 2019

[11] Sathiyaprasad, B., Seetharaman, K., & Kumar, B. S. (2020). Content based Video Retrieval using Improved Gray Level Co-Occurrence Matrix with Region-based Pre Convoluted Neural Network–RPCNN. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 558-563). IEEE.

[12] Dai C, Liu X, Yang LT, Ni M, Ma Z, Zhang Q, Jamal Deen M (2020) Video scene segmentation using tensor-train faster-rcnn for multimedia iot systems. IEEE Internet of Things Journal

[13] Golgiyaz S, Talu MF, Onat C (2019) Artificial neural network regression model to predict flue gas temperature and emissions with the spectral norm of flame image. Fuel 255:115827

[14] Fablet Ronan, Patrick Bouthemy and Patrick Pérez, “Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval”, IEEE Transactions on Image Processing vol. 11, no. 4, pp. 393-407, 2013.

[15] Ferman A. Müfit A. Murat Tekalp and Rajiv Mehrotra, “Robust color histogram descriptors for video segment retrieval and identification”, IEEE Transactions on Image Processing, vol. 11, no. 5, pp. 497-508, 2014.

[16] Sahoo P. K., P. Kanungo, and S. Mishra,"A fast valley-based segmentation for detection of slowly moving objects, Signal, Image and Video Processing, vol. 12, no.7, pp. 1265-1272, 2018.

[17] Roy Partha Pratim, Ayan Kumar Bhunia, and Umapada Pal, "Date-field retrieval in scene image and video frames using text enhancement and shape coding", Neurocomputing, vol. 274, pp. 37-49, 2018.

[18] Mühling M., Korfhage N., Müller E., Otto C., Springstein M., Langelage T. and Freisleben, B, "Deep learning for content-based video retrieval in film and television production", Multimedia Tools and Applications, vol. 76, no. 21, pp. 22169-22194, 2017.

[19] Kumar, B. S., &Seetharaman, K. (2021). Video sequence feature extraction and segmentation using likelihood regression model. Multimedia Tools and Applications, 1-19.