International Journal on Emerging Technologies **10**(3): 384-390(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Image and Video Frame Extraction System Based on Improved Deep Learning Technique

S.N. Sithi Shamila[1], D.S. Mahendran[2] and M. Mohamed Sathik[3]
[1]Part-time Research Scholar (Registration no. : 12474)
Assistant Professor, Department of Computer Science, Manomaniam Sundaranar University, Abhishekapatti,Tirunelveli-627 012, (Tamil Nadu), India.
[2]Principal, Aditanar College of Arts and Science,Tiruchendur, (Tamil Nadu), India.
[3]Principal,Sadakathullah Appa College,Tirunelveli, (Tamil Nadu), India.

(Corresponding author: S.N. Sithi Shamila)
(Received 15 July 2019, Revised 28 September 2019 Accepted 04 October 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Nowadays deep learning technique is widely used in image processing. It is highly responsible for recent growth in the use of artificial intelligence. Deep learning technique is used in this proposed work because most of the data which is present will be in unstructured format because most of it exists in different format such as text, audio, video and PDF files. Unstructured data is difficult to analyze for most machine learning algorithms. In such cases deep learning can be used and it can be trained by using different data formats. The target of this paper is to achieve high accuracy in image and video extraction. In this paper, we have discussed how image processing works, along with deep learning technology. The feature which is extracted from the video is in the form of text, image and audio. It is capable of retrieving desired videos by selecting relevant one and filtering out undesired videos. It makes predictions by using informations in a dataset and uses that experience in real world. Because of its high accuracy, retrieval of image and video extraction is done by using Convolutional Neural Network (CNN) and Fisher vector method.

Keywords: Convolutional neural network method (CNN), fisher vector method, Gaussian filter and Median filter

Abbreviations: EHD, Edge Histogram Descriptor (EHD); CNN, Convolutional Neural Network; EMD, Earth Movers Distance.

I. INTRODUCTION

In our daily life, Image processing plays an important role in the fields of science and technology. It is a computer based technology. It carries out automatic processing, manipulation and interpretation of visual information which is retrieved from the video [6], [5]. In this huge world, very complex problems get solved by using these kinds of techniques. Video data possesses a lot of information for those using multimedia systems and applications like digital libraries, publications, education, broadcasting and entertainment. Video retrieval is done based on content. It focuses more on textures, images and color [1, 3]. For retrieving videos and images, cloud based computing framework [7], big data and image queries are used but it is not effective [2, 8, 13]. Several techniques were used for retrieval of image and video. While extracting, the date may be in structured and unstructured format. Unstructured data cannot be processed easily by most machine learning algorithms. This proposed paper focuses mainly on processing the unstructured data and to get the high quality image. For concatenation of partial derivatives and to elucidate in which direction the parameter of the model should be modified to best fit the data, Fisher vector method is used. Convolutional Neural Network is most widely used for image related problems. It is used for image classification and recognition because of its high accuracy. The main part of convolution model is convolutional layer. The main part of convolutional neural network is convolution layer. A computer understands an image using numbers at each pixel. CNN compares the feature one by one and after the comparison, chosen feature was put it on the input image if it matches then the image is classified correctly. After the classification of image, Lines up the feature and the image then multiply each image pixel by the corresponding feature pixel and the final step in convolutional layer is to add them up and divide by total number of pixels in the feature to get the output. The next layer is ReLU layer, it removes the negative values from the filtered image and replaces it with zero's. This is done to avoid the values from summing up to zero and the third step is to perform pooling operation. It calculates the largest value of all feature map and the results are down sampled and finally all the feature map is to be shrunk to get a 4 × 4 matrix. While extracting the main purpose of feature extraction is to decide whether there is any noise or not. If we found any noise it should be removed by using filters such as Gaussian filter and median filter. After removal of noise, the next step is to detect the shapes from extracted feature and in upcoming steps merging of information is performed. While comparing to its predecessors. The main advantage of CNN is, it has best level performance on problems that significantly outperforms other solutions in multiple domains such as speech, languages etc, without any manual intervention it automatically extract the significant features. CNN is great in capturing single features in the bottom layers of an image as well as complex features and entities in the deeper level [17]. It breaks the images down into numbers and then merges multiple set of information pooling them together to

create an accurate representation of images. A large amount of data can be used effectively within a deep learning model.

In section II, video retrieval based on deep learning process is discussed; section III shows the result evaluation. Image and video recognition based on deep learning is discussed in section IV, finally problems and challenges are discussed in section V, VI shows Result with the conclusion in section VII.

II. VIDEO RETRIEVAL BASED ON DEEP LEARNING PROCESSES

A. Capturing a video

A video is captured by using a high resolution camera. A captured video is to be processed further to provide a high accuracy.

B. Segmentation of Video

The first step used after capturing a video is segmentation of video into shots. These shots contain a sequence of features which is taken one after another to form a video event. It breaks images down into numbers and then merges multiple set of information and pooling them together to create an accurate representation of an image. A large amount of data can be used effectively with a deep learning model.

C. Noise Removal

While extracting the feature, noise exists. It should be removed by using techniques such as Gaussian and median filter

Gaussian Filter: Gaussian filter blur the images by using Gaussian function. It is widely used to remove image noise.

Median filter: The median filter is a non linear digital filtering technique used to remove salt and pepper noise and to sharpen the edges.

The pre processing step is used to improve the result of image processing.

D. CNN based video extraction

Due to large number of pixels and high dimensionality and working with large quantities of digital images becomes difficult. So for extracting such type of features, a technique called convolutional neural network was used. This model is referred to as the unsupervised convolutional Siamese Network. Pre – trained CNN modules are adopted to extract visual features from intermediate layer of convolution. By using aggregation function, these features are computed through the forward propagation of an image over CNN network.

E. Filtering of extracted feature

The feature which is extracted from the video is in the form of text, image and audio. It is capable of retrieving desired videos by sharply selecting relevant one and filtering out undesired videos which exist. It makes predictions by using information in a dataset and uses that experience in real world.

F. Features and Features Extraction from video

While extracting the feature, three primary features are extracted. They are color, texture and motion. These are represented by color histograms and motion histograms. While extraction, the most important features that are to be included in the video are features of the objects, key frames and the motion features. Key frame feature in video contains color, texture and shape. These are the most important features of visual properties. RGB, HSV, YCbCr and normalized r-g, YUV, and HVC are the extracted color features. They play an important role in video indexing and retrieval. Various Techniques are used for finding energy distribution in frequency domain while extracting the texture features. While extracting the features there exists some noise before segmentation and filtering and those noises should be removed by using Gaussian and median filter. A co-occurrence matrix is a matrix or distribution of co-occurring values of an image. It represents texture in images. The matrix elements are the counts of the number of times a given feature occurs in a particular spatial relation to another given feature. A co-occurrence matrix can use any of the features from the image. GLCM is the co-occurrence matrix when grey level is selected as a feature. The GLCM tabulation shows the different combinations of pixel grey levels occur in an image. An example to find GLCM of a matrix of Fig. 1 having grey values 0, 1, 2, 3 are shown here and its GLCM is shown in Fig. 2.

0	0	1	1
0	0	1	1
0	2	2	2
2	2	3	3

Fig. 1. Matrix.

Reference pixel values \ Neighbour pixel values	0	1	2	3
0	2	2	1	0
1	0	2	0	0
2	0	0	3	1
3	0	0	0	1

Fig. 2. GLCM of the matrix of Fig. 1.

For effective video retrieval, textual features can be utilized. Textual feature play a vital role in video retrieval. A texture feature vector is generated by using mean and variance of the filtered outputs. The image is split into small blocks and it is used to obtain features from these blocks. It divides the image into blocks with each of size 5 x 5 and compute texture features from each block. Features and the shapes of the objects are extracted from the edge and by using histograms regional features are extracted. Various shape descriptors are utilized but the shape descriptors used here are Edge Histogram Descriptor. EHD is one of the widely used methods for shape detection.

An Edge Histogram Descriptor (EHD) is designed by dividing an image into 4 x 4 blocks. The partition of image creates 16 equal sized blocks. The spatial distribution of edges is obtained and then, categorized into five different orientations of 0, 45, 90, 135 degrees.

The EHD is the number of pixels forming an edge of a particular category. Local Histogram alone is not applicable for high retrieval performance so beside the local histogram, global histogram is also needed. The global edge histogram represents the edge distribution of whole image space. In Edge Histogram Descriptor, initially the RGB image is converted into grey image and then divides the image into 4 × 4 blocks. After these steps percentage of number of pixels that correspond to an edge histogram is computed and also the global edge histogram is computed by using same procedure. Finally save both the local and global histogram values in feature vector.

Motion features: The two types of motion feature are background feature and foreground feature. By moving objects the background feature is created and by camera motion the foreground feature is generated. The motion feature caused by camera motion includes zooming in or zooming out, panning left or right and tilting up or down by camera. The motion features caused by object plays an important role in describing motions of key objects.

Extraction of Gabor features: Gabor features contains a group of wavelets with each wavelet capturing energy at a specific frequency and a specific direction. The filter which is used in Gabor is designed to detect different frequencies and orientations. From each filtered image, Gabor features can be calculated and used to retrieve images. The algorithm for extracting the Gabor feature vector is shown in Fig. 3 and the related Eqns. (1-4) are shown below. For a given image I (x, y), the discrete Gabor wavelet transform is given by a convolution:

$$W_{mn} = \sum_{x1}\sum_{y1} I(x1, y1) g_{mn} * (x - x1, y - y1) \quad (1)$$

where m and n indicates the scale and orientations of wavelet respectively. After applying Gabor filters on the image with different orientation at different scale, an array of magnitudes is obtained:

$$E(m,n) = \sum_x \sum_y |W_{mn}(x, y)| \quad (2)$$

In different scale and orientation of the image, the magnitude represents the energy content. The main purpose of texture-based retrieval is to identify the images or regions with relevant texture.

The standard deviation σ of the magnitude of the transformed coefficients is:

$$\sigma_{mn} = \sqrt{\frac{\sum_x \sum_y (|W_{mn}(x,y)| - \mu_{mn})^2}{PXQ}} \quad (3)$$

where μ is the mean of magnitude and given as

$$\mu_{mn} = \frac{E(m,n)}{PXQ}$$

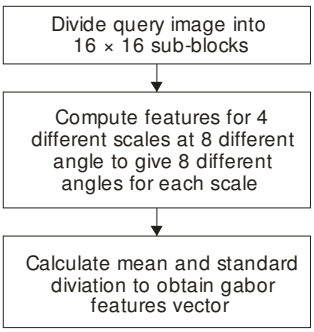

Fig. 3. Gabor Filter Algorithm.

A feature vector f (texture representation) is created using mn as the feature components. M scales and N orientations are used and the feature vector is given in equation

$$f = [\sigma_{00}, \sigma_{01}, \sigma_{02}, \ldots . \sigma_{(M-1)(N-1)}] \quad (4)$$

$f_{Gabor} = \frac{f-\mu}{\sigma}$ where μ is the mean and σ is the standard deviation of f.

G. Similarity Measure

According to the type of feature, Queries are classified. The query is identified by computing similarity between the feature vectors which are stored in the database. Obtaining the similarity with the acquired still image which is extracted from the video clip is measured from the database with the help of similarity between the feature vectors through a distance between them. To identify the similarity between the feature vectors, Euclidean distance is measured. Image similarity is performed by measuring the distance between the query image and the image which is in the database. Performance of video retrieval system is determined from the type of features. Once feature is created, enhancement of performance is done to get better results from similarity measure. Euclidean distance and Minkowski type distance is widely used for similarity. Similarity is performed between the query video and the video which is in the database. Matching is done with the extracted features, texts, objects and faces with the video or image which is in the database. At each different levels of resolution, video similarity can be measured. A video clip is captured by identifying the key frames that are occurred continuously in the video database which is related to the query video. The distance metric is termed as similarity measure. To rank the retrieved videos, the Euclidean distance between the query and database is calculated. The video from the database corresponding to the frame similar to the query frame is higher in rank if the Euclidean distance is smaller. The equation for Euclidean distance between the query image Q and an image P is shown in Eqn. (5)

$$ED = \sum_{i=1}^{n} \sqrt{(V_{pi} - V_{qi}).(V_{pi} - V_{qi})} \quad (5)$$

where, V_{pi} and V_{qi} are the feature vectors of Query image Q and image P respectively of size 'n'. To measure the distance between the two features, there are several methods to measure other than Euclidean distance are Earth Movers Distance (EMD), chord distance etc. To measure the distance between two distinct probability distributions, Kull back and Libeler method is used. The equation for KL divergence of the probability distributions F, G on a finite set P is given in Eqn. (6).

$$D_{KL}(F||G) = \sum_{p\varepsilon p} F(p) log \frac{F(p)}{G(p)} \quad (6)$$

Below are the steps for Similarity Measure: Let us consider - F as Query clip feature vector, G as Feature library 1st feature vector, *i* as Element of vector, M as Normalized factor of G

$$V = \frac{F}{\text{Normalization}(F)} \quad (7)$$

Then find ((G>0) & (V> 0)) and store that in VA. Then similarity measure is computed by using Eqn. (8)

$$D_{KL} = \sum V(V_A) log \frac{M * V(V_A)}{G(V_A)} \quad (8)$$

Neural Network can also be utilized to identify the similar shots. Clustering of shots is performed to classify videos to the best matching cluster based on the features which is extracted from its shots. In object based query. The features of color and texture in a shot are used to map the shot to the best matching cluster. Similarity between the query image and an image which is found in the video database is obtained by probability of generating the image.

III. RESULT EVALUATION

The performance of video retrieval is computed with the same parameters as it is evaluated in retrieval of image. Recall and precision are the two parameters as given in Eqns. (9) and (10).

Recall = DC/DB (9)

Precision = DC/DT (10)

DC = number of similar clips detected correctly

DB = number of similar clips in the data base

DT = total number of detected clips

IV. IMAGE AND VIDEO RECOGNITION BASED ON IMPROVED DEEP LEARNING

Deep learning technique is widely used nowadays for the process of image recognition and video retrieval. In deep learning, convolutional neural network is applied for analyzing the visual imagery. It is a multilayer perceptron. The "fully connected" of the network is susceptible for over fitting of data. Fisher vector method combines the advantage of statistical models and discriminative methods. For concatenation of partial derivatives and to elucidate in which direction the parameter of the model should be modified to best fit the data, Fisher vector method is used. CNN uses little preprocessing step when comparing to other image classification algorithm. In CNN, detection of video, image classification and similarity of image query with the image in the database is performed. CNN is specialized in processing the data. The significant part of convolutional neural network is convolution layer. In Convolution layer, A computer understands an image using numbers at each pixel. CNN compares the feature one by one and after the comparison, chosen feature was put it on the input image if it matches then the image is classified correctly. After the classification of image, Line up the feature and the image then multiply each image pixel by the corresponding feature pixel and the final step in convolutional layer is to add them up and divide by total number of pixels in the feature to get the output. Detection of video is performed initially. After detection process, removal of noise is performed by using Gaussian and median filter. Gaussian filter is used to blur the image. It is used to reduce the noises from the input image. To remove the noise and to get the enhanced image, the gray scale image is passed through Gaussian filter and median filter. Median filter is used to remove the salt and pepper noise. If the noise is removed, the further processing can be done in an efficient way to achieve high performance. After the removal of noise, the next layer is to be performed is to detect the shapes from the extracted feature and then image segmentation is performed. Image segmentation is done by breaking down images into numbers and finally merging of information is performed to create an accurate representation of image. A large amount of data can be used effectively with a deep learning model. The main advantage of CNN is that without any manual intervention automatic extraction feature is performed. CNN is best in extracting the features in the bottom layer of image as well as complex feature and entities in the deeper level. Finally similarity measure is performed by comparing the image query with the image which is stored in the database. Euclidean distance is measured between feature vectors. A video clip is captured by identifying key frames which is relevant to that of the query video. A video component i.e., shots, scenes and frames etc are extracted from the videos and then classified to predefined categories. Extraction of feature from each component and is stored in the feature database. The output video is obtained by finding the similarity measure between features of query video and features stored in the database.

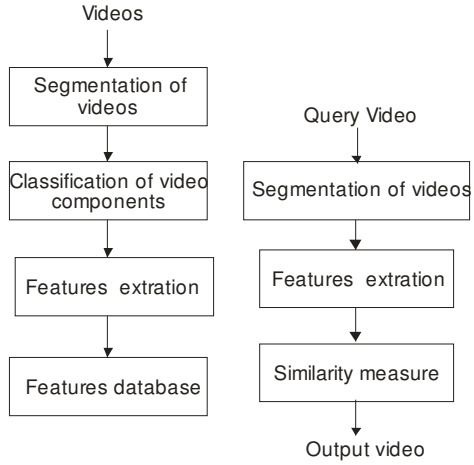

Fig. 4. CBVR system.

To enhance the performance of retrieval system, relevance feedback technique is used to resemble human visual judgment and similarity perception to a certain extent. Relevance feedback technique is very useful in obtaining effective ranking and retrieving similar videos. It removes the difference between low level features and semantic concept of the videos. It relies on feedback obtained by user or can be automatic and accordingly the videos are ranked. To improve the features, ranking and feedback is used. Relevance feedback is used in the system [2]. The result is obtained by updating the values of Mu and updating of Mu is done by method shown below.

$$u = x, y$$

The weights M_x and M_y are updated using user's feedback. Let S be the set containing the most similar L retrieved video clips, overall similarity value Hv and value of Mx and My is 0.5

$S = [S_1, S_2, \ldots\ldots, S_L]$

$Score = [Score_1, Score_2, \ldots\ldots Score_L]$

be the set containing scores by relevance feedback by the user for each retrieved clips in set S. The scores may have any of the values from -3, -1, 0, +1, +3. Where these values correspond to the feedback as

+3 → highly relevant

+1 → relevant
0 → no opinion
-1 → non-relevant
-3 → highly non-relevant

Mx and My are the sets containing the most similar L clips to the query, according to only the color similarity measure and only the motion similarity measure, respectively.

$S^x=[S_1^x, S_2^x,\ldots,S_L^x]$
$S^c=[S_1^c, S_2^c,\ldots,S_L^c]$

Weights of Mu are updated using the value of score provided by the user as a feedback. Weights of Mu are more for the more relevant retrieved clips. The weights are then normalized by the total weights to make sum of the normalized weight equal to 1 and if the weight of Mu is < 0, then it is set to 0. The system can be iterated to improve the result for a satisfaction level. As a result, a particular feature representation will represent the semantic concept of the query video.

V. PROBLEMS AND CHALLENGES

The feature which is extracted cannot be processed directly and some modifications have to be done to get high accuracy. There are difficulties in classifying images and segmentations. It causes a big problem for security and validation purposes. CNN is great in capturing single features in the bottom layers of an image as well as complex features and entities in the deeper level. But there is difficulty in representing the rotational and translational relationships. So CNN is to be explicitly trained very well. CNN is good for recognizing the different elements of a face such as eyes, nose and mouth but it is not more sensitive to the arrangement of the entities and made mistake in perceiving different arrangement of the entities. So CNNs needs to be managed to handle translational invariance.

VI. RESULT

Feature method	True Positive	False Negative	Accuracy %
Surf	480	61	88.72
Freak	470	71	86.87
Surf-Freak	489	52	90.38
Surf-Freak Optimized	507	34	93.71

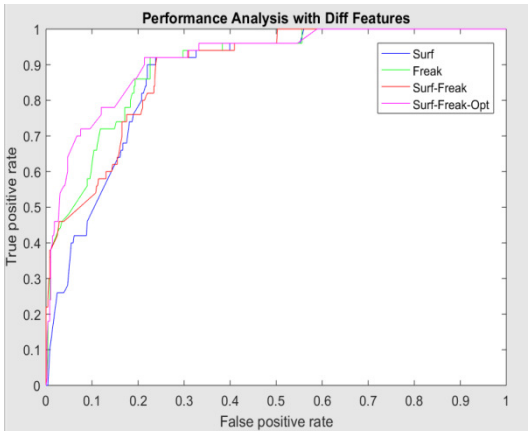

Fig. 5. Performance analysis with different features.

From the above Fig., it is clear that the optimized Surf-Freak based feature extraction performs well from the initial time. The proposed feature extractor used more number of features to get accurate result. For content based image retrieval the above features can be used individually but for the video retrieving system a hybrid combination of features as proposed above are needed to get better result.

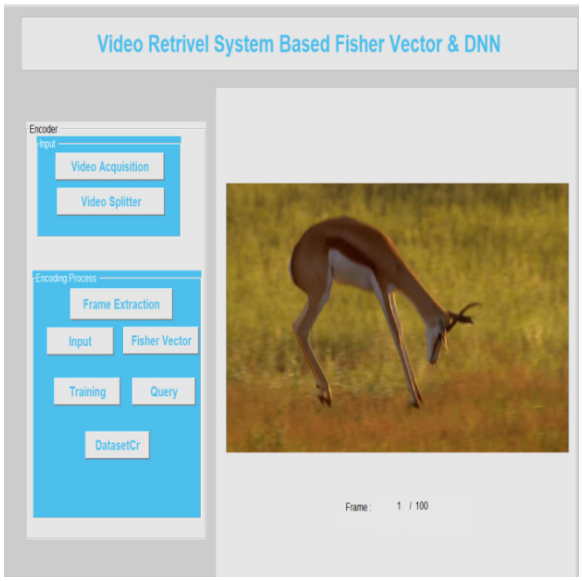

Fig. 6. A sample video is taken.

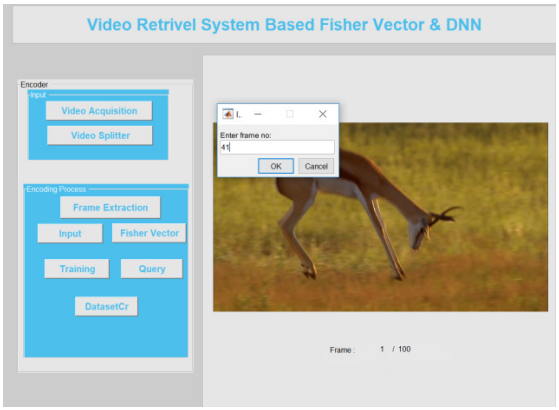

Fig. 7. Specific frame is selected.

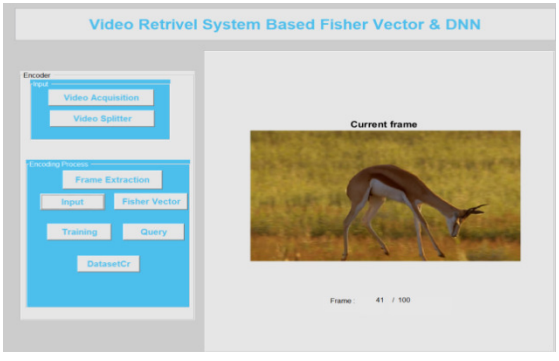

Fig. 8. Selected frame is fed to system.

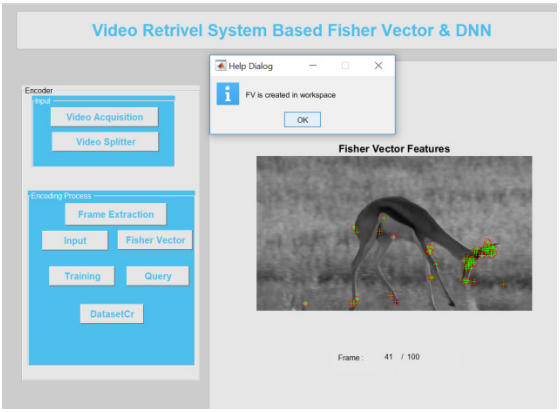

Fig. 9. Features selection process.

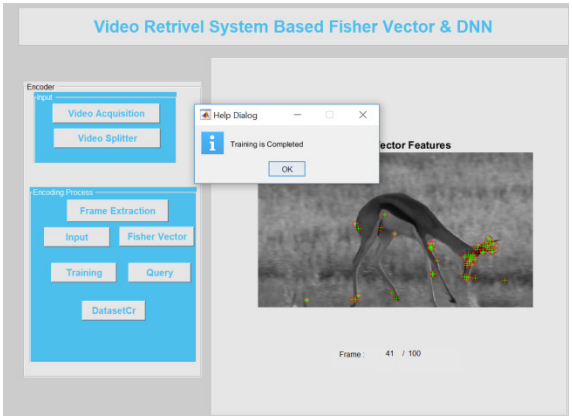

Fig. 10. System is trained with proposed features.

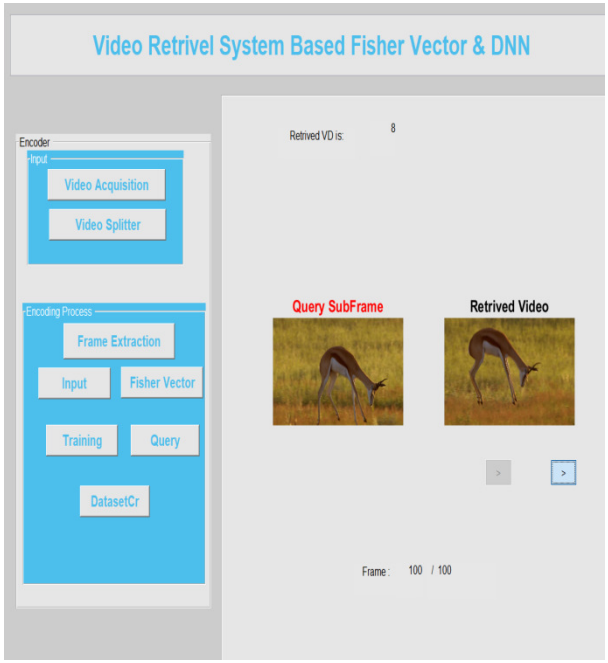

Fig. 11. Output Video sequence.

VII. CONCLUSION

In this paper, Convolutional Neural Network is most widely used technique for image related problems. While comparing to other works the main advantage of CNN is, It has best level performance on problems that significantly outperforms other solutions in multiple domains such as speech, languages etc. Convolutional neural network is computationally efficient. It has a great advantage in identifying a model of unstructured data which cannot be done by other machine learning algorithms. For better performance, it uses special convolution and pooling method. After performing the convolution operation, pooling method is performed to reduce the dimensionality and it reduces the number of parameters which shortens time consumption for training data and combats over fitting. It is very powerful technique to deal with efficient models which extract features to achieve better accuracy.

REFERENCES

[1]. Yang, H., & Meinel, C. (2014) Content Based Lecture Video Retrieval Using Speech and Video Text Information. *IEEE Transactions on Learning Technologies,* Vol. *7,* pp. 142-154.

[2]. Araujo, A., & Girod, B. (2018). Large-Scale Video Retrieval Using Image Queries. *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. *28, pp.* 1406-1420.

[3]. Jing-Ming Guo, Heri Prasetyo, Jen-Ho Chen 2015, "Content-Based Image Retrieval Using Error Diffusion Block Truncation Coding Features. *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. *25*(3), pp. 466-481.

[4]. Chien-Li Chou, Hua-Tsung Chen, Suh-Yin Lee (2015). Pattern-Based Near-Duplicate Video Retrieval and Localization on Web-Scale Videos. *IEEE Transactions on Multimedia,* Vol. *17*(3), pp. 382-395.

[5]. Lokoc, J., Bailer, W., Schoeffmann, K., Muenzer, B., & Awad, G. 2018. On Influential Trends in Interactive Video Retrieval. *IEEE Transactions on Multimedia,* Vol. *20*(12)*, pp. 3361-3376.*

[6]. Zhang, L., Wang, L., Lin, W., & Yan, S. (2014). Geometric Optimum Experimental Design for Collaborative Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. *24*(2), *pp. 346-359.*

[7]. Hanli Wang, Bo Xiao, Lei Wang, Fengkuangtian Zhu, Yu-Gang Jiang, Jun Wu (2015). A Cloud-Based Heterogeneous Computing Framework for Large-Scale Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. *25*(12), pp. 1900-1913.

[8] Yinan Feng, Pan Zhou, Jie Xu, Shouling Ji, Dapeng Wu 2019, "Video Big Data Retrieval Over Media Cloud: A Context-Aware Online Learning Approach. *IEEE Transactions on Multimedia*, Vol. *21*(7), pp. 1762-1777.

[9] Yifang Yin, Yi Yu, Roger Zimmermann 2015, "On Generating Content-Oriented Geo Features for Sensor-Rich Outdoor Video Search", IEEE Transactions on Multimedia, Vol. *17*, no .10, pp. 1760-1772.

[10]. Jianfeng Dong, Xirong Li, Cees G. M. Snoek 2018, "Predicting Visual Features From Text for Image and Video Caption Retrieval", IEEE Transactions on Multimedia, Vol. *20*, no. 12, pp. 3317-3388.

[11]. Seshadri Padmanabha Venkatagiri, Mun Choon Chan, Wei Tsang Ooi, Jia Han Chiam (2015). On

Demand Retrieval of Crowd sourced Mobile Video", *IEEE Sensors Journal, Vol. 15*, no. 5, pp. 2632-2642.
[12] Maurizio Montagnuolo, Paolo Platter, Alessio Bosca, Nicolo Bidotti, Alberto Messina (2019). "Real time Semantic Enrichment of Video Streams in the Age of Big Data", *SMPTE Motion Imaging Journal, Vol. 128,* no. 2, pp. 1-8.
[13] H. Yang, B. Quehl, and H. Sack 2012, "A framework for improved video text detection and recognition", Multimedia Tools Application, pp. 1–29
[14] Gregory Castanon, Mohamed Elgharib, Venkatesh Saligrama, Pierre-Marc Jodoin 2016, "Retrieval in Long-Surveillance Videos Using User-Described Motion and Object Attributes", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. *26*, no. 12, pp. 2313-2327
[15] Chongke Bi, Ye Yuan, JiaWan Zhang, Yun Shi, Yiqing Xiang, Yuehuan Wang, RongHui Zhang 2018, "Dynamic Mode Decomposition Based Video Shot Detection", *IEEE Access,* Vol. *6*, pp. 21398-21407
[16] S. Tippaya, S. Sitjongsataporn, T. Tan, M. Khan, and K. Chamnongthai 2017, "Multi-modal visual features-based video shot boundary detection", *IEEE Access, Vol. 5*, pp. 12563–12575
[17] Cesc Chunseong Park, Youngjin Kim, Gunhee Kim 2019, "Retrieval of Sentence Sequences for an Image Stream via Coherence Recurrent Convolutional Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. *40,* no. 4, pp. 945-957.
[18] Yuan-Hao Lai, Chuan-Kai Yang 2015, "Video Object Retrieval by Trajectory and Appearance", *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. *25,* no. 5, pp. 1026-1037.
[19] Lin Ding, Yonghong Tian, Hongfei Fan, Yaowei Wang, Tiejun Huang 2017, "Rate-Performance-Loss Optimization for Inter-Frame Deep Feature Coding From Video", *IEEE Transactions on Image Processing, Vol. 26, no.12, pp. 5743-5757*.
[20] Gengshen Wu, Jungong Han, Yuchen Guo, Li Liu, Guiguang Ding, Qiang Ni, Ling Shao (2019). "Unsupervised Deep Video Hashing via Balanced Code for Large-Scale Video Retrieval", *IEEE Transactions on Image Processing,* Vol. *28*, no. 4, pp. 1993-2007.

How to cite this article: Shamila, S.N.S. Mahendran, D.S. and Sathik, M. (2019). Image and Video Frame Extraction System Based on Improved Deep Learning Technique. *International Journal on Emerging Technologies*, **10**(3): 384–390.