International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 09 | Sept-2020 ISSN: 2582-3930

Content Based Video Recommendation of Surveillance Videos

Paul Francis, Shreya Shalom Jo, S JyothisYadu, Sneha Anna Thomas [1], Dr.Jina Varghese [2]

[1]BTECH Computer Science and Engineering, Amal Jyothi College of Engineering
[2]Assistant Professor at Department of Computer Science and Engineering, Amal Jyothi College of Engineering

---

Abstract -Mostly videos are retrieved from video archives, in response to user queries. The system replies with a series of roughly relevant videos by matching the input text with the title related to the video. However, the content of the video might not be indistinguishable with the input text given bythe user. In the proposed system, the contents of the video are processed to seek out the specified video clip, in line with the user requirement. Given a query from the user regarding the presence of an interested person, the system retrieves relevant videos by matching the input keyword with the content of the video. The major attraction of the proposed scheme is the fusion of summarization technique to reduce the cost of video processing. The content matching is performed insummarized videos. Accordingly, face recognition and video summarization are the major parts of the proposed system. This method has an added advantage when deploying in huge organizations with many numbers of survelliance videos. The proposed system will produce better accuracy compared to the literature work. The proposed system is applicable within the enforcement department, broadcasting system, etc. This technique makes CCTV footage retrieval easier and within lesstime**.**

Key Words:Key Frame Extraction (KFE), face recognition, CCTV video, Convultional Neural Network (CNN)

1.INTRODUCTION

Before discussing about video processing and extraction, let us discuss about what video is and about itscharacteristics and constraints. A video may be a sequence of 2 dimensional images, that are captured into a plane by digital devices like video camera from a dynamic 3 dimensional scene. The colours in an image of a video frame are the reflected light at a3Dpointwithinthescenefromwhichthesceneiscaptured. Videos are often transmitted or transported via medium like wireless terrestrial television, transmission line etc.Video could also be transmitted over networks and other digital platforms.

Each video may vary counting on its different attributes like display resolution, frames per second, aspect ratio, bit rate and many other qualities. Let us discuss few of such attributesthatplayataskinourproject:

- **Number of frames per second**: Frames are the still images within the video. Frame rate is that the number of frames per unit time of video. It is measured in frames/s with FPS as its unit. Standard frame ratesmay range from 25 – 30 frames/s. At least 16 FPS isrequired todevelopaclosefittedmovingimage.
- **Aspect ratio**: It is the ratio between width and height ofvideoscreenorelementsofthevideo.
- **Bit rate**:Bit rate may be a measure of the rate of data content of the digital video stream. Uncompressed videos exhibits maximum quality, but also has high bit rate . Bit rate is an important characteristic that determines the standard of the video since itis proportional to properties that affect the quality of video. Bit rate is a pivotal property when it comes to video transmission because the transmission link must be able to support that bit rate. It also relevant when the storage of the video file in concerned since the video size and its duration is dependent on the bit rate. Video optimization techniques can be employed to reduce the bit rate to an extent while having a minor effect on quality.

Discussed below are the various parameters used for various applications of video

Application	Frame rate	Dimensions	Pixel Depth
Multimedia	15	320 x 240	16
Entertainment TV	25	640 x 480	16
Surveillance	5	640 x 480	12
Video Telephony	10	320 x 240	12
HDTV	25	1920 x 1080	24

TABLE I: Application of digital videos and their constraints

With top quality and size there arises the matter of storage constraints. Although the frame rate of surveillance video is the least as mentioned in Table , the dimensions may be a bigger issue since CCTV videos are bulky. For the better processing of such cumbersome files, it's crucial to scale back the dimensions and data.

Size is a very important aspect when it involves storage and transmission. There exist different methods of multimedia storage like drives, memory cards, flash disks to call a couple of . The trending and most generally being cloud storage. Physical devices accompany the impediment of storage capacity. Especially while considering substantial number of videos. With the growing technologies and population there comes the necessity of huge storage, which may be accessed at any time and place, thus endorsing the virtual memory services. Cloud storage is that the most ordinarily used virtual platform which enables its worldwide users to store copious data. Both virtual and physical storage have their own setbacks. Physical storage devices could also be susceptible to data loss upon damage, storage constraints, andcan'tbeaccessedreadilybutvirtualstoragealsofaces the matter of security and network issues. If the network requirements for cloud storage are often met, it's the most recommended option for storage purposes.

Before moving further, let us spot the termvideo processing and the way it works. Video processing could also be a specific case of image processing, where the output and input signals are video files or video streams. Video processors are used to wield apparent definition of video signals.Theywilldothesubsequenttasksonavideo:

- deinterlacing
- aspect ratiocontrol
- digital zoom andpan
- brightness/contrast/hue/saturation/sharpness/gamma adjustments
- frame rate conversion andinverse-telecine
- color pointconversion
- color spaceconversion
- mosquito noisereduction
- block noisereduction
- detailenhancement
- edgeenhancement
- motioncompensation
- primary and secondary colorcalibration

The image processing and video processingtechniques are pertinent in motion and object detection. Thus, having most of its application in traffic management so on the analyze the video sequences in traffic flow, traffic data collection and road traffic monitoring.After the advancement of latest display devices, the technological aspect of video processing gets broader. With the emergence of latest media and network, the transmission and storage inherit being. so on transmit such rich multimedia content, it's desirable to occupy less bandwidth, the size of the primary video signal must be reduced by some compression technique, without degrading video quality or data loss. to urge obviate suchhurdles, video compression techniques provide efficient solutions to represent video data during a more compact and robust manner so as that the storage and transmission problems with video are often realized in cost effectiveway.

A. Overview

The evolution of multimedia data types and available bandwidth the demand for video retrieval systems is additionally growing because the users deviates from text based retrieval systems to content based retrieval systems.Content-Based Video Retrieval (CBVR) system assists an admin to retrieve sequence (target)from anout sized database where it might be difficult for the operator to do a manual search . Selection of extracted features plays vital role in content based video retrieval no matter video attributes."Content-based"searchwillanalyzetheparticular content of thevideo.

In this modern era, CCTV cameras have an inevitable role in recognizing the victims and offenders of an illegal activity. It has become an inevitable part of our law enforcement. The process is complex since it operates on huge volume of data for finding a particular shot/scene from these large videos.

For example, in an organization, the data stored from CCTV might be huge on a daily basis depending on the number of surveillance cameras. The data of a CCTV camera is 1GB per day which makes the storage management a tedious task. To find the sequences containing the face of a person would take days of manual work,since the image is being stored in such a big database,as the amount of work done on database for searching a face increases.Such a system cannot quickly retrieve items when needed.This might results in the slow retrieval of data.It takes a long time to find the information about a relevantperson[2].

Over the past decade, there are disparate efforts in solving this problem. There exist copious works to detect objects, faces or movements in CCTV videos. But with growing interests there comes the necessity to enhance the efficiency and time frame of the output. Therefore, we are proposing a content-based video retrieval system for CCTV videos.

2. RELATED WORKS

The need for content-based access to image andvideo information from media archives has captured the attention of almost all researchers in recent years. Research efforts have led to the development of methods for Computer Vision and Pattern Recognition. The methods are used to determine the similarity in the visual information content extracted

from low level features. In 2012, [3] proposed the concept of "Content based video retrieval systems" through their research paper, which discusses the video segmentation technique and different feature extractions like audio,image,shots,colorof objects,facesetc that may be used for CBVR systems in near future. In 2017, [6] designed and validated a "Smart surveillance based on video summarization". It carries out the smart summarization of surveillance videos using the approach of video summarization. They use feature extraction to produce summarized frames. Similarity of the frames and measured and appropriate video is retrieved after comparison. Furthermore, in 2016 [Byeon et al.(2016)Byeon, Pan,Moh, and Kwak] developed a surveillance system Using CNN for face Recognition with object, human and face detection. The system was able to detect objects and people within 2-5 m distance and identify them with an accuracy of 72.2%. But the efficiency of the system is questioned when it comes to inability of the system to identify certain people even after training with 500 images of each person. In 2009, [7] presented a framework and a data model for CBVR system for CCTV surveillance videos on RDBMS providing the functions of a surveillance monitoring system, with a tagging structure for event detection. In 2016,[5] conducted a review on CBVR and analysis using image processing. The paper discusses on various works related to CBVRandexperimentsonthesame.

Eventhough,thereexistsanumberofworksrelatedtoface recognition and CBVR those poses the dispute of efficiency and time. In some, with video optmization efficiency is lost and in other time and data needed to produce an efficient system is relatively high. Hence, we propose a system that uses Key frame extraction as well as an efficient CNNbased face recognition system for surveillance videos using OpenCV inPython.

3. PROPOSED SYSTEM

The proposed system aims to retrieve the clipsfrom input videos which includes the queried person of interest. The system has to identify the person in the video toretrieve the respective shot or clip. To identify the person, a training of the persons should be done in prior. This is done by Convolutional Neural Network. The proposed system utilized LBP haar cascades for it. Fig 1shows the framework of the proposedsystem.

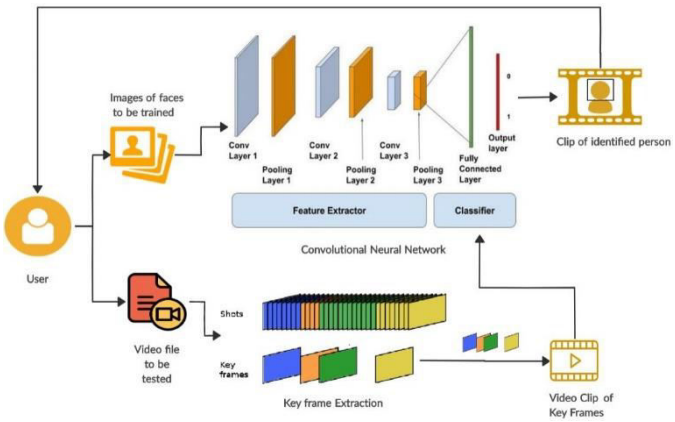

Fig. 1: Framework of proposed system

A. *KEY FRAMEEXTRACTION(KFE)*

Key frame extraction is a major step in reducing the cost of the video retrieval systems. Multimedia contents are huge in terms of the points which needs processing. The general structure of a video is given in the Fig. 2. The normal display rate of a video is 25-30 fps(frames per second). The information conveyed in a video clip of one second duration is meaningless. Processing all these 30 frames to check the presence of an interesting object or human is costly. The key frameextractiontechniqueshasanimportantrolehere.

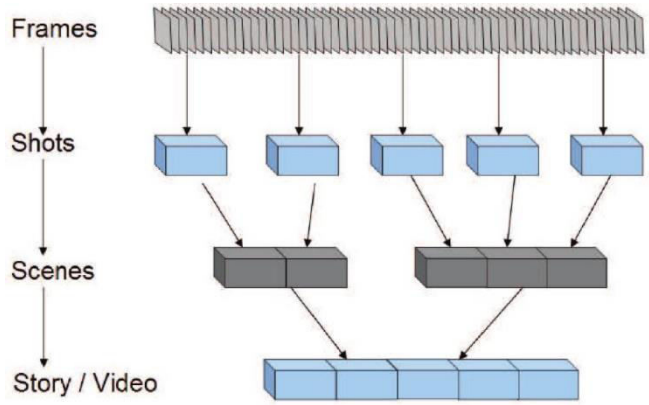

Fig. 2: Key frame extraction from video frames

1) Extraction of relevant frames: The three types of frames in video were analysed and the *P* frames convey major informations than *B* and *I* frames. In thecompression of the video stream, frames can be grouped into sequences called a group of pictures (GOP). The types of frames can be classified into I-frames, P-frames and B-frames. They are regularly arranged in the video stream and compose the GOPs. An I-frame is a complete image and does not contain motion vector displacements. P-frame holds only thechanges in the image from the previous frame. It contains bothimage data as well as motion vector displacements. We use P-frame extraction as it is more efficient than I frame extraction. In I-frame extractionwe lose the details in the contents of the

videos whereas in P-frame extraction we are able to retrieve frames the without losing any details of the video along with reducing repetitive frames.This set of keyframes is produced with no loss in semantic quality and continutiy of the video. An enduser never experiences a distortions in the summary video, which makes the approach appreciable.

The extracted key frames from a movie is shown in the fig. 4 The frames in the original video are shown in the fig. 3. Major representative frames from the input video are picked as the keyframes as illustred in fig, 4and fig. 3. The analysis of key frame extraction technique is detailed in the *RESULTS AND ANALYSIS* section.

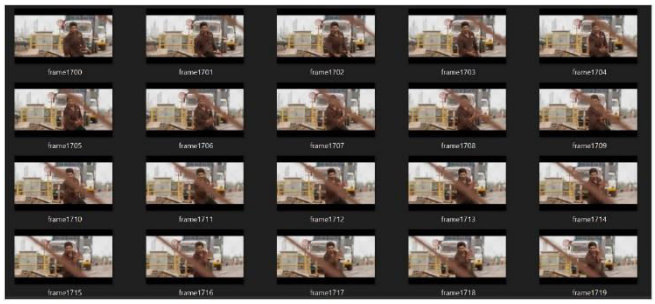

Fig. 3: Frames in Original Video

Fig. 4: Extracted Key frame

2) FACERECOGNITION

Face recognition is a tedious task. Face recognition plays an inevitable role in real world applications such as video surveillance, human machine interaction and security systems . Deep learning based methods have shown better performances in terms of accuracy and speed of processing in image recognition as compared to machine learning approaches.This method has a rapid classification approach which requires normalization and some pre-processing to exhibit better classification performance than other approaches such asEigenfaces. The reason is the reduced number of training database(which is sometimes varied from1 - 5).

*1) FaceModelTraining:*Fortraining,multiplesnapsof every person should be supplied since the face recognition results depends completely on the face images collected for training set.Snaps taken should contain different expression postures including both frontal and non-frontal face taking into account characteristics such as light conditions, movements, device quality etc. Most of the challenges faced on dealing with face recognition is because of the predicament that faces are not rigid objects and pictures are oftentakenfrommanyvariousviewpointsoftheface.

*a) LBPclassifieres:*LocalBinaryPattern(LBP)isan efficient texture operator used to label the pixels of an image. It is done so by thresholding the neighborhood of each pixel and considering the result as a binary number. It is a very popular mechanism in numerous fields due to its computational simplicity and discriminative power, thus bringing together the statistical and structural models of texture analysis. Its robustness and computational simplicity makes it more suitable for real-world applications. The overall working of LBP classifier is depicted in fig.5[4].

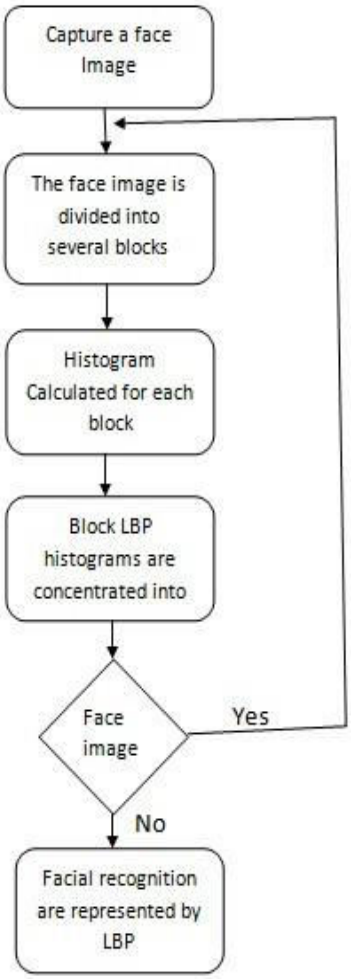

Fig. 5: LBP Classifier Working

*b) Training::*WithLBPitispossibletodescribethe texture and shape of a digital image. This is done by dividing an image into several small regions from which the features are extracted.These features consist of binary patterns that describe the surroundings of pixels in the regions. The obtained features from the regions are concatenated into a single feature histogram, which forms a representation of the image. Because of the way the texture and shape of imagesis described, the method seems to be quite robust against face images with different facial expressions, different lightening conditions, image rotation and aging of persons. Once the Local Binary Pattern for every pixel is calculated, the feature vector of

the image can be constructed . For an efficient representation of the face, first the image is dividedinto K^2regions.Infigure1.7afaceimageisdividedinto 82 = 64 regions. For every region a histogram with all possible labels is constructed. This means that every bin in a histogram represents a pattern and contains the number of its appearance in the region. The feature vector is then constructed by concatenating the regional histograms to onebighistogram.

2) *Cataloging Persons in the Input Video:* Training ofthe facesisdonewithLBPclassifiersandCNN.Thistrained model can then be used to detect and identify the persons in the given input video. Parameters of the training is given in the TableII.

Category	Number of images	Image Size
Face1	50	240 x 120
Face2	50	240 x 120
Face3	50	240 x 120

TABLE II: Parameters of training

Fig.6displays the result of face recognition testing done on a home video. It can be seen that the face are identified and labeled with the name of the person given whiletraining. Along with this, the confidence percentage is alsodisplayed.

Fig. 6: Face recognised

B. OUTPUTGENERATION

The trained model is enforced in the given input video to catch the presence of interested face. A tabular output is prepared to respond to the user queries like is theperson A present in the video? Who all are present in this video? How many faces? When does this person arrive in this video? the location of this CCTV footage?. The clips of the interested faces are thenextracted.

4. RESULT AND ANALYSIS

A. DATASET

For *training* purpose a set of images having variant resolutions and features were chosen. These includes frontal and non-frontal faces, different hairstyles, with and without accessories like spectacles, hats etc. For *testing* the model, mainly home videos were used.

B. KEY FRAME EXTRACTION

For analysing the result of KFE the following parameters areconsidered:

- **Number of frames** : Used to display the number of framestheinputvideohasbeforereduction.
- **Number of key frames** : The number of key frames thatarededuced,fromthesetofallframes,afterKFE
- **Reduction percentage** : It shows the percentage of reduction that took effect. Eqn.1 can be used to find the reduction percentage where *VBR* is the Value before reduction and *VAR* is the value afterreduction.

$$Reduction\ Percentage = \frac{VBR - VAR}{VBR} \times 100 \quad (1)$$

No.of frames	No. of key frames	Reduction %	Video Duration	Duration after KFE	Reduction %
3470	1440	58.50%	2.4 min	1.6 min	33.34%
11,087	9662	12.85%	6.16 min	5.36 min	12.98%
12,543	9750	22.26%	8.05 min	6.2 min	22.98%

TABLE III: Reduction percentage of frames and time

C. FACE RECOGNITION

Face recognition model parameters:

1) Training: Training was performed for 2 faces and the number of samples being 168 and 34 photos of each person. The training data sets were chosen such a way that each person's sample photos includes frontal and non-frontalfacesand are of differentresolutions.

2) Testing in Real time videos: A set of 5 home videos were chosen for testing. The parameters like recall(sensitivity/hit ratio), accuracy, precision and F1-score can be calculated with the help of confusion matrix. Let us consider that our model is designed to identify a person 'X'. Then the outcomes of the confusion matrix for that model can be describedas:

- **True Positive (TP)** : identifies 'X' as'X'
- **False Positive (FP)** : identifies another person as'X'
- **False Negative (FN)** : identifies 'X' as anotherperson
- **True Negative (TN)** : identifies some one else as 'NOT X'

The following equations can be used to calculate the parameters:

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$F_1 - Score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (5)$$

Category	Recall	Accuracy	Precision	F1 Score
Video1	100%	98.4%	90.9%	0.952
Video2	75%	90.9%	81.8%	0.783
Video3	90%	87.2%	69.2%	0.783
Video4	100%	80.8%	61.5%	0.762
Video5	85.71%	70%	54.5%	0.667

TABLE IV: Result assessment of face recognition

D. ANALYSIS OF THE RESULT

Table III shows the details regarding the videos that endured reduction. Reduction percentage in terms of number of frames and the duration of the video is determined using Eq. (1). It can be observed that after KFE the frames are reduced by an average of 31.26% and average timereduction percentage is 23.1%. Fig.7and fig.8 are the graphical analysis of the same. It can be seen that the KFE needsmuch more improvement than the existing version which can beachievedinthenearfuture.

The videos after key frame extraction and face recognition isanalysed and the performance is recorded as shown in Table IV. It can be observed that the system performs pretty well with an average accuracy of 85.46%, recall of 90.142%, precisionof71.58%andagoodF1scoreof0.7894.

5. CONCLUSION AND FUTURE WORKS

With the event of multimedia data types and available bandwidth there's huge demand of video retrieval systems, as users shift from text based retrieval systems to content based retrieval systems.The main contribution of this paper is to get a strong content based video with high accuracy. In this paper, we use a CNN architecture along with feature extraction technique.

The general structure of face recognition process in this paper is formed from pre-processing stage that contains collecting images,making sub directories for every character that continues with extraction of facial expression, and afterwards extracted feature set is assessed . In our system, KFE is employed to perform segmentation of video into elementary shots. CBVR has steps as key frame extraction, feature vector formation, similarity and template matching and eventually get the retrieved relatively correct expected copiesofvideosapproximatelymatchingwithqueryvideo. Invideos,sizable amount of frames forms a scene.Thisscene includes repetition of nearly same frames with slight differences,whichincreasesthespaceforstoringandreduces the performance in video processing. Here rather than searching the entire set of frames in videos, only selected key frame are used for further processing. Key frames are extractedbycomputingtheconsecutiveframedifferences.

For future works, we can improve the KFE to enhance the reduction further more and thus improve the performance, accuracy and speed of our system by reducing data to be tested.In the near future,the have the following advancements:

- Detect facial expressions andemotions
- Improve key frameextraction
- Performtherecognitionofdamagedordeformedfaces
- Detect the suspiciousactivities
- Enable the option of virtual storage for storing large number of videos and implement any time any whereaccessibility
- Performdetectiononrealtimevideostreams

Simultaneous testing of multiplevideos

REFERENCES

1. Y.-H. Byeon, S.-B. Pan, S.-M. Moh, and K.-C. Kwak, "A surveillance system usingcnnforface recognition with object, human and face detection,"in*InformationScience and Applications (ICISA) 2016*. Springer, 2016,pp.975–984.
2. T. S. Jebara, "3d pose estimation and normalizationforfacerecognition," *Centre for Intelligent Machines, McGillUniversity*,1995.
3. B. Patel andB. Meshram, "Content basedvideo retrieval systems," *arXiv preprint arXiv:1205.1641*, 2012.
4. M.Rahim,M. Hossain, T. Wahid, and M. Azam, "Face recognition using local binary patterns (lbp)," 01 2013.
5. S. Srinivasa Raghava, Y. Janarthanan, and J. Balajee, "Content based video retrieval and analysis using image processing: A review," *International Journal of Pharmacy and Technology*, vol. 8, no. 4, pp. 5042–5048, 2016.
6. S. S. Thomas,S. Gupta, and V. K. Subramanian, "Smart surveillance based on video summarization," in *2017 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 2017, pp. 1–5.
7. Y. Yang, B. C. Lovell, andF. Dadgostar, "Content-based video retrieval (cbvr) system for cctvsurveillancevideos,"in*2009DigitalImageComputing:Techniquesand Applications*, 2009, pp.183–187.