ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

Article Title: Surveillance Video Summarization based on Target Object Detection

Surveillance Video Summarization based on Target Object Detection

Dr. D. Minola Davids[1], Dr. C. Seldev Christopher[2], Dinesh B.G. Wilson[3]

[1]AP, ECE Department, C S I Institute of Technology, Thovalai, Nagercoil, Tamilnadu, India, dminoladavids@yahoo.co.in

[2]Professor, CSE Department, St. Xavier's Catholic College of Engineering, Chunkankadai, Nagercoil, Tamilnadu, India, cseldev@gmail.comD

[3]AP, Department of Mechanical Engineering, CSI Institute of Technology Thovalai, Nagercoil,Tamilnadu, India dinesh.wilson @gmail.com

ABSTRACT

The recent trend increases the use of surveillance cameras in many of the private and public premises, which causes the number of surveillance videos to grow exponentially. The information gained from these surveillance videos not only helps the owner of the property, but also helps in crime investigations for police and security officials. Though there are several applications of such videos, yet their storage, management and retrieval processes are still challenging. Hence, it is important to develop an efficient technique to describe a long video into a shorter video with semantic information by eliminating the redundant and unimportant frames. This technique makes the larger video to shrink in length for efficient storage and also helps the users to attain a complete knowledge of the video by only watching the shorter video, without spending more time in watching the original longer video. To achieve this objective, this paper proposes a video summarization technique for summarizing the surveillance videos by extracting the target object using YOLO then discarding the remaining frames and finally combining the extracted key frames into a single video. This method first detects the target object related in the original video frames and then eliminates the remaining frames that are irrelevant without prominent objects, resulting in video with only the key frames which are into the interest of the user, finally those frames are combined to form a summarized video.

Keywords: Video Summarization, Surveillance Videos, Target Object Detection, Key Frame Extraction.

1 Introduction

In current trends, security has become the primary concern. Surveillance systems, especially the surveillance cameras play a vital part in security field. This makes the use of CCTV cameras in larger number for the past decade. They are installed in both private and public premises. The video outcomes of these cameras are going on increasing in a daily basis. Hence, analyzing, managing, storing and retrieving such videos are found to be challenging. Such videos could be more helpful for several purposes other than the owners, including the

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

police, security officers, crime investigation specialists for their investigation activities as major evidences.

To get the required or specific information in a surveillance video, the user should watch the complete video, which is time consuming and uninteresting. The screen time spent should be more for watching a full video, which leads the user to miss some important information in it. Hence, these videos should be processed for useful information based on the user's interest. Video processing for extracting such video contents is challenging. This is because, a single surveillance video has a large number of redundant and unessential frames. This motivates the researchers to develop an efficient and effective way of analyzing the surveillance videos to extract the key frames based on the user interest, though preserving the video integrity completely.

Several techniques are available for performing video summarization, in which many are dealt with entertainment, sports, etc. The outcomes of those techniques would be based on the highlights of the sport activities in sports videos, trailers of the videos of entertainment and so on. Hence, it is clear that there are several ways to achieve video summarization. Basically, the video summarization means the highlights in a original video or a short summary of a longer video. This summarized video should consist of highly important features, whereas eliminating the redundant features.

This paper is organized with a brief description of motivation and evolution of video summarization in section-2, the related works that are studied has been described in section-3, proposed methodology is demonstrated in section-4 followed by experimental results and discussion in section-5 and finally this paper is concluded in section-6.

2 Video Summarization

Video summarization means the process of making a long video into a shorter summary by selecting only the important or highly informative frames and combining them into a single video. The summarized video generally consists of a set of key frames that are of special interests, extracted from the original longer video. The main objective of video summarization is that, to make the browsing of full video content faster and to efficiently access the fames of interest. Hence, decision making can be quicker and easier regarding the utility of the video and for any suspicious actions. The process of selecting the key frames will be varied based on the target users or the applications to which the video is going to be helpful. This can be identified by performing usability studies for measuring the quality of the video summary.

The popularity of video capturing devices and advancement in web technologies over the past years the video data are increasing dramatically, thereby demanding some efficient and effective tools for browsing, accessing and manipulating large videos. Thus, speeding-up a video is the easier way to browse quickly the contents in the video. This can be done by either re-encoding or reducing the frames of a video. It is possible to reduce the video

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

browsing time by using the fast playback option. This method can be applicable nearly without any pitch distortion with the help of a technology called time compression. This technique seems to be easier and can preserver many useful information, so that the video watching time can be reduced. But, the lack of comprehensive knowledge about the original video, the frame selection will be impossible thereby most of the redundant frames were still available in the shortened video. Hence, further limiting the video duration will not be possible. Moreover, the there is a limitation of ratio of time compression.

Analysis of videos are in research since 1990 and several approaches were introduced for structuring and understanding the contents of a video automatically. This induces the method of video summarization to select the key frames with most of the information with more comprehension of the contents in video. Video summarization relates to the process of video skimming and video abstraction. These techniques will be differed in their outcomes. In video skimming, short clips of the video are skimmed from original video to form a new short video with informative materials. In video abstraction and video summary, a set of key frames or clips or other media like texts are created to provide a short description of the original video. Currently, video summarization plays a major role in providing fast indexing and browsing of a larger video or collection of videos.

3 Related Works

This section provides a brief study of existing works carried on in the field of video summarization. A method of motion detection-based video summarization which makes use of the combination of background subtraction and Structure-Texture-Noise Decomposition to handle the sensor noise and variations in illumination in a scene. In this method, the gray-level sequences are decomposed into structure, texture and noise components, in which the structure and texture components are extracted using Aujol decomposition, which can then be used for background modeling. This background is subtracted from the binary image to detect any objects moving. Before detecting the motion, the noise should be removed. A traditional cosine measure based-alarm similarity module is proposed to minimize the computational complexity in the process of motion detection [1].Video summarization based on deterministic indexing and selection of key frames is demonstrated in [2]. This method is analyzed using 2 datasets namely CAVIAR, CViSOR. After completely extracting the frames from a video, the Otsu method is used in computing the global difference using Adaptive Threshold Setting. The histogram features are extracted from each frame, so as to reduce the computation time and to preserve the 2D information in the frames. The key frames are then selected based on the histogram correlation. Finally, the extracted key frames are clustered and aligned to create a summary. Video summarization using Convolutional Neural Network (CNN) for surveillance videos taken from resource constrained devices has been described in [3].In this method, three steps are followed viz., shot segmenting based on deep features which is extracted from the fully connected layer of CNN trained on Mobile Net, then

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

computation of image memorability and entropy and then selecting the key frames from the shots and finally the selected shots are summarized. The duplicate frames are discarded using the difference in color histogram.

Frame re-composition-based video synopsis approach is discovered in [4]. This method uses a video cube of dimensions (X, Y, T). Here, T is the length. The spatio-temporal trajectories are extracted by using a combination of background subtractor, dense optical flow and clustering which helps in forming the spatio-temporal tublets. These tublets are deployed individually to create a video summary through reducing the space between the 3D structures. An optimized video summarization framework technique is described in [5] which converts the content based-video retrieval into image retrieval. In this method, an optimal framework is developed for perceptual video summarization, which summarizes the video based on human perception to extract the key frames and again an optimal framework is developed for perceptual video retrieval, which extracts a frame as a single background for the extracted frames that are stitched as a single video shot. Video summarization by extracting the keyframes by maximizing the diversity and representational ability is demonstrated in [6]. This method follows two phases. The first phase is the video partition, in which the surveillance video is partitioned into visually coherent video chunks having more inter chunk variation and high intra-chunk resemblance. This method is applied for measuring the significance between 2 frames and performs video partition by using Normalized Cut algorithm. In the second phase, the partitioned video chunk is subjected check for frames having the most attention score and are chosen as the attentive frame using the method of content completeness along with visual satisfaction.

A context-aware video summarization technique which uses the important information in the video frames, has been explained in [7]. To collect the frames with important information, by extracting the local motion regions and the interactions among them. This method follows sparse coding with comprehensive sparse group lasso to study the video feature dictionary and spatio-temporal feature correlation graph dictionary. The importance of the information in the extracted features are ensured by sparsity. A joint approach of video summary based on images and video captioning based on text has been described in [8]. This method uses an encoder-decoder framework which is a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) and the outcome of this approach is the video summary along with its captions. This model is evaluated using two datasets namely the MST-Video to Text dataset and Microsoft Video Description Corpus. In [9], it is defined that the intensity and nature of actions can be applied for summarizing the video. This means that the frames detected for any motion is called actionness and these frames are chosen and summarized. This method initially begins the temporal actionness concept and its relationship with video summarization. Then, a set of actionness labels were introduced over four

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

summarization benchmarks. They are analyzed for consistency verification. At last, temporal actionness is used to enhance the video summary via multitask learning.

In [10], a new technique for finding the first person who wears the camera via a third person's camera. Here, the actions such as object tracking, activity recognition and joint scene understanding have been carried out on the videos captured by the cameras worn by the third person in their body to identify the first person. This can be done by measuring the distance between the first and the third person from their videos, which is a tedious process, since the first person will not be visible in the video captured from his/her camera. For this approach, a semi-Siamese CNN is used to solve the above-mentioned limitation. This method learns the low-level features individually for first and third person videos, while sharing the high-level features and embedding space for measuring the distance. Video summarization based on Attentive Encoder-Decoder framework is provided in [11], which is based on the attention mechanism, i.e., it does not consider all the frames as same, instead providing weights for each frame based on their importance. This framework is based on supervised learning making use of Long Short-Term Memory (LSTM) to develop the methods such as additive and multiplicative attention mechanisms. Both the methods obtained better performance on the datasets like TV Sum and Sum ME for additive and multiplicative attention mechanisms respectively. Video summarization of multi-video based on hypergraph dominant set is demonstrated in [12]. It is a method of clustering to meet multi-video summarization by forming it as a problem of discovering common visual patterns. The keyframes are the centroid of dominant sets. A method called query dependent maximum marginal relevance has been presented for making the summary query adaptive. This technique balances the minimal redundancy and query adaptive criteria to create a set of key frames. Also, a method called graph based topical closeness has been presented to improve the logic and legibility of the video summary generated.

Most of the existing techniques creates video summarization informative for a single video. But in multiple videos for informative summarization generation, they got failed. Hence, in [13], a Multi-modal Weighted Archetypal Analysis for generating an informative and representative video summarization from multi-video has been proposed. This method fuses the information of the frames, web-based images and tags through multi-modal graph. Hence, this system is based on graph and multi-modal fusion approach. Also, this system exploits Archetypal analysis algorithm for creating summarization based on query. Hence, it is based on a decomposition approach. As given in [13], another one method of multi-video summarization has been demonstrated in [14], using an unsupervised framework. An approach called diversity-aware sparse optimization has been used in this method for summarizing the video. The problem of optimization has been solved efficiently by using alternating minimization algorithm. This algorithm limits the general objective function of a video, at the time of setting other videos. An unsupervised method of video summarization

has been implemented in [15], in which a sparse subset is selected from the video frames that optimally characterize the video given as the input. In this method, the keyframes with minimum distance between the feature representation and the video. This distance is specified using Generative Adversarial Network (GAN) to optimize selecting the frame. A variational auto encoder LSTM is used as the summarizer in this approach. Another one LSTM is used as discriminator for differentiating the original and the reconstructed videos.

A multi-view representative selection optimization for video summarization using multi-view sparse dictionary selection with centroid co-regulation is proposed in [16]. This method identifies the consensus choice of visual features that are view-specific. The final video summary was observed to have different visual elements covered, with the help of diversity regularizer. With the incorporation of some view-specific selection priors, the quality of the summarized video had been improved using external data and supervision. The method in [17] is implemented for summarizing surveillance video of multiple views. The single-view video summarization techniques are the traditional way of video summarization, which cannot be successfully applied for summarizing multi-view video summarization, since the complications in inter-view and intra-view correlations, cannot be able to exploited completely. Both these correlations are exploited in a joint embedding space. This can be obtained by solving a problem of an Eigen-value which is linear in the number of multi-view videos. Sparse representative selection is employed over the embedding space for video summarization. To achieve the multi-view video summarization, the same authors in [17], proposed another one approach based on an unsupervised framework which is given in [18], which uses joint embedding and sparse representative selection. In this method, the multi-view correlations are captured at first, which are then used to extract diverse set of representatives. Then $\ell_{2,1}$-norm is used for modeling the sparsity at the time of representative shots selection. A half-quadratic minimization algorithm is used for convergence analysis.

The unpaired data can be used for video summarization. With a raw video as input, a set of frames with some key features are identified and are combined to form a meaningful video summary. Most of the current techniques on video summarization are based on supervised learning. This type of approach needs a large amount of labeled data for training purpose. Manual annotation by humans is also required for raw videos and ground truth summary videos. Hence, the training process is highly expensive and hard to create the training samples with labels. To overcome the above limitation, summarizing the video using unpaired data can be helpful. This method is not expensive and easy due to fact that large number of raw videos and good video summaries can be readily available in internet and can be easily accessible [19]. In videos with high frequency of viewing, tracking the temporal structures and enforcing local diversity is necessary. The local diversity denotes that the selected frames should be diverse in short duration, but the frames can be the same if they occur lately. A sequential determinantal point process based probabilistic model is proposed

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

in [20]. It controls the video time span dynamically by executing local diversity. This model learns to automatically conclude how local the local diversity has been found in a video given as input. Model training is performed using reinforcement learning algorithm, which solves the problem of non-differentiable evaluation measures and exposure bias occurred due to the use of Maximum Likely-hood Estimation (MLE) in model training.

Based on the background study of several video summarization techniques, video summarization using interest-based object detection using ResNet model based on Convolutional Neural Network is found to be more effective and efficient, since in this method, the key frames that are non-redundant and with semantic information based on surveillance purposes are only considered for the video summarization, which makes the final summarized video with useful contents related to surveillance systems to help the owners, police, crime investigation department and security professionals.

4 Proposed Methodlolgy

This section provides the proposed methodology of surveillance video summarization technique. The proposed video summarization is based on detecting the target object using YOLO (You Only Look Once) v3 and discarding the frames without prominent objects and finally summarizes the extracted key frames with target objects. This summarized video is the outcome of the proposed system. The block diagram of the proposed system is given in Figure 1. The modules involved in this method are briefed as follows:

1. *Input video selection:* The input of the proposed video summarization model is an original surveillance video obtained from a CCTV camera.
2. *Target object detection:* The video frames with the target objects are detected and then the remaining frames without any target objects are all discarded. Now only the key frames with the target objects are available.
3. *Video summarization:* Finally, the all the key frames with target objects are combined to form a summarized video.

These models are explained in detail as follows:

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

Article Title: Surveillance Video Summarization based on Target Object Detection

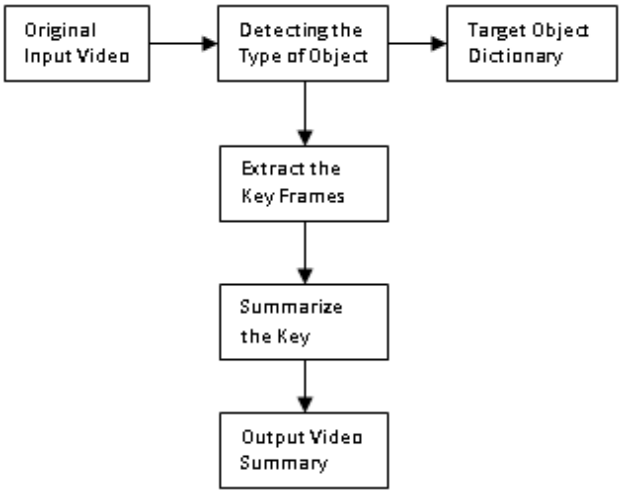

Figure 1: *Proposed Video Summarization Technique*

4.1 Input Video

The proposed system develops a desktop application with an interactive and user-friendly user interface using Python. The input raw surveillance video is fed into the YOLO model, which then extract the target objects for extracting key video frames. The video format supported in the developed application are the MP4 and AVI.

4.2 Selecting the Target Objects

Once the input video is selected, then the target objects are needed to be selected, in order to extract the frames of key interest, i.e., surveillance systems. For this, the MS COCO dataset is used to develop a dictionary. This dataset is composed of 330,000 images, out of which has 200,000 labeled images. There are 80 categories of objects with 15,000,000 instances. The objects include, person, indoor and outdoor objects, food, kitchenware, electronics, vehicles, sports etc.

4.3 Detecting the Target Objects

As already mentioned, that target object detection is carried out using YOLOv3. The target objects are searched over the scenes, events and frames for locating it. A Darknet variant is used in YOLOv3. It has 53 layers trained on Imagenet. Moreover, another 53 layers are added for efficient object detection. This makes the architecture of YOLOv3, a fully convolution with totally 106 layers. The convolutional layer in this architecture uses a stride 2 for down-sampling the feature maps to avoid low-level feature loss. It has no pooling layer. A single neural network is applied on a full video to split the frames into regions, also boundary boxes and probabilities will be predicted. The Figure 2 shows the architecture of YOLOv3 and Figure 3 represents the prediction of bounding boxes.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

Article Title: Surveillance Video Summarization based on Target Object Detection

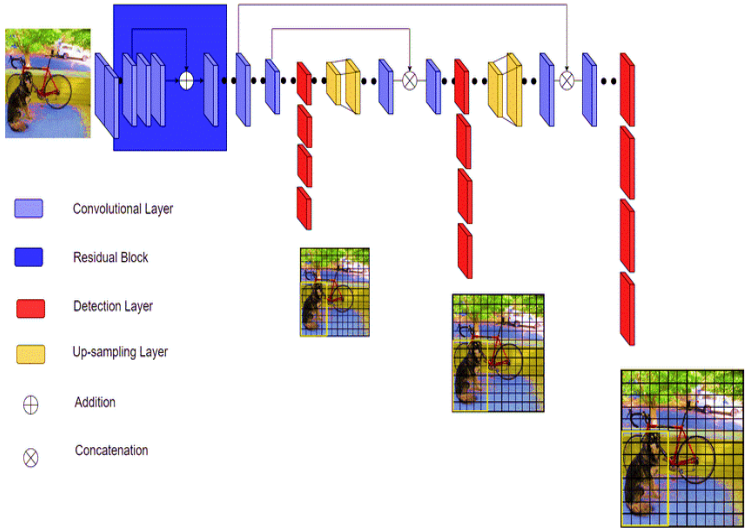

Figure 2: *Architecture of YOLOv3*

Logistic regression approach is used to predict the scores of each class and a threshold is used to predict the objects with multiple labels. The class with scores more than the threshold will be bounded by a box, which represents the target object detected. If a frame contains many objects, then spatial location of the target object can be defined using this method.

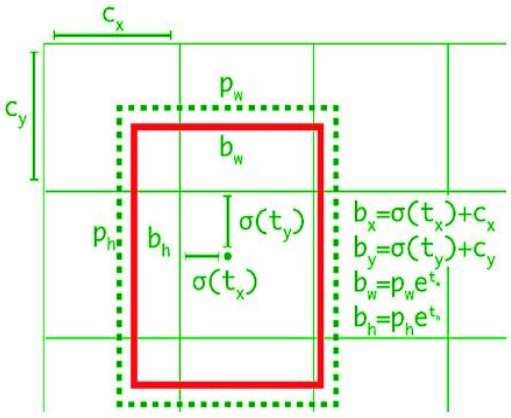

Figure 3: *Prediction of Bounding Boxes*

The reason behind the use of YOLOv3 in the proposed approach for video summarization is that, the YOLOv3 is a faster approach for object detection, when compared to the other algorithms. The comparison of speed of YOLOv3 with other algorithms are described in Table-I and graphically represented in Figure 5, and the comparison of accuracy of YOLOv3 with other algorithms are described in Table-II and graphically represented in Figure 6. The YOLOv3 has a processing speed of 45fps which is higher than that of the Single Shot

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

Detector (SSD), Region-based Fully Convolutional Network (R-FCN) and Faster-Region based Convolutional Neural Network (F-RCNN).

4.4 Video Summarization

The video summarization is performed by combining all the frames with target objects as the input to the module and provides the output with a summary of the original surveillance video. The steps involved in the summarization of the surveillance video are explained as follows:

At first, the current frame of the input surveillance video has been read, then detection of target objects has been performed in that frame using YOLOv3 and if found, then that frame will be saved in the buffer, else discard that frame. Continue this process for all the frames one by one. If the frame currently processing is the last, then summarize the saved frames with target objects in the buffer, by combining all of them. The algorithm of proposed surveillance video summarization technique is given in Algorithm 1.

Algorithm: Generation of Video Summary

Input: Surveillance video (I), Target object frames (J)

Output: Video Summary (S)

$$VS(I,J)$$

$$n \leftarrow Number\ of\ frames\ (I)$$

for $i = 0$ to $n - 1$, **do**

Read current frame $C[i]$

Status of Target Object detection $(F[i], J)$

if$(status == 1)$**then**

{

$S[i] = F[i]$

}

else

Discard frame

end for

Save S

---

5 Experimental Results and Discussion

The relevance of the proposed method has been demonstrated by experiments, which has been described in this section. The proposed system is implemented on Jupyter notebook with Python programming language. The input of this system is the surveillance video obtained

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

from a CCTV camera outside of a house. This video is divided into frames and are fed into the YOLOv3 algorithm for detecting target objects trained using MS COCO dataset. For instance, classes including vehicles, animals and humans are considered as the target objects. Hence, the YOLOv3 will detect those objects and draws a boundary box over it. These frames are extracted and are saved into the buffer for further processing. Among these frames, only the unique frames are selected for summarization process and all the redundant frames are discarded. All the unique frames with detected target objects are finally combined together to form a summarized video output of the input surveillance video. The Figure 4 shows the detection of target objects like dog, bicycle and truck using YOLOv3 and MS COCO dataset in a video frame of a CCTV camera footage obtained from a house.

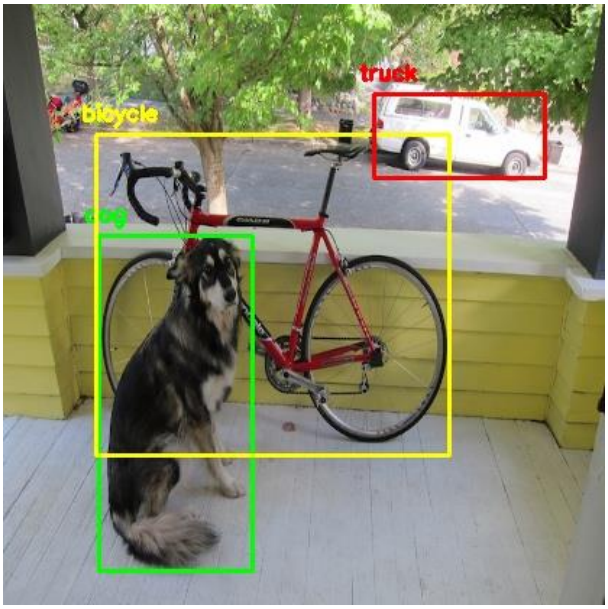

Figure 4: *Target objects detected in a video frame using YOLOv3*

6 Performance Evaluation

The performance of the proposed surveillance video summarization approach is evaluated using the performance evaluation metrics such as the Accuracy, Precision, Recall and F1-Score, which are described and expressed mathematically as follows:

6.1 Accuracy

The classification accuracy is a spontaneous tactic which is defined as the ratio of sum of predicted values to the sum of corrected predicted values and is represented mathematically as,

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100 \quad (1)$$

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

Article Title: Surveillance Video Summarization based on Target Object Detection

6.2 Precision

The precision is defined as the ratio of true positive predicted values to the sum of total positive predicted values and is represented mathematically as,

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

6.3 Recall

The Recall is defined as the ratio of true positive predicted values to the sum of true positive and false negative predicted values and is represented mathematically as,

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

6.4 F1-Score

The F-Measure (F_β) is defined as the weighted average or the harmonic mean of Precision and Recalland is represented mathematically as,

$$F_\beta = \frac{(1+\beta^2)(Precision*Recall)}{(\beta^2*(Precision+Recall))} \tag{4}$$

Substitute the value of β as 1, then the F_1 score can be represented as,

$$F_1 = \frac{2*(Precision*Recall)}{1*Precision+Recall} \tag{5}$$

Table 1: *Comparison of the speed of YOLOv3 and other algorithms*

Algorithm	Processing Speed (fps)
Single Shot Detector (SSD)	22
Region-based Fully Convolutional Network (R-FCN)	6
Faster-Region based Convolutional Neural Network (F-RCNN)	17
Proposed YOLOv3	45

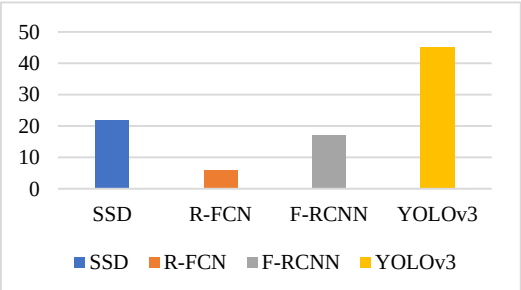

Figure 5: *Graphical representation of speed comparison of YOLOv3 and other algorithms*

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

Article Title: Surveillance Video Summarization based on Target Object Detection

Table 2: *Comparison of the accuracy of YOLOv3 and other algorithms*

Algorithm	Accuracy (%)
Single Shot Detector (SSD)	26.8
Region-based Fully Convolutional Network (R-FCN)	31.5
Faster-Region based Convolutional Neural Network (F-RCNN)	21.9
Proposed YOLOv3	33

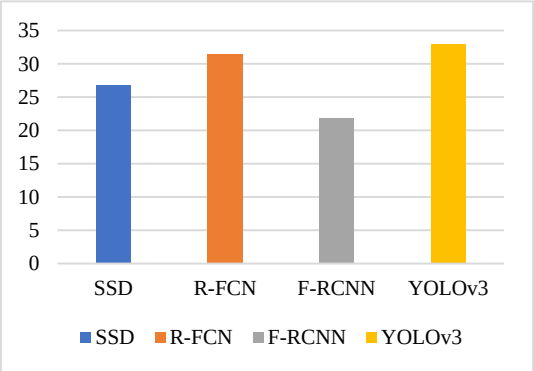

Figure 6: *Graphical representation of accuracy comparison of YOLOv3 and other algorithms*

7 Concusion

An efficient and effective way of video summarization technique has been described and demonstrated in this paper. This video summarization framework was developed for summarizing a long surveillance video using target object detection. The proposed video summarization method performs well in terms of processing speed and accuracy. The use of frames with target object for video summary, makes the proposed system less complicated, more optimal flexible and reliable. The use of YOLOv3 for target object detection in the proposed approach empowers the system for detecting various objects with more efficiency and precise. MS COCO dataset is used for model training with several classes, which makes the proposed system capable of detecting almost all of the objects in real-world. Only the frames with target objects and are unique will be summarized for output video summary and the frames without target objects and are redundant will be discarded. The proposed framework for surveillance video summarization is verified to generate a very short, precise and accurate video summary.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
PP. 852-866

Article Title: Surveillance Video Summarization based on Target Object Detection

References

1. Omar Elharrouss; Noor Al-Maadeed; Somaya Al-Maadeed, Year: 2019, "Video Summarization based on Motion Detection for Surveillance Systems", IEEE 2019.

2. T Michael Moses; K Balachandran, Year: 2019, "A Deterministic Key-Frame Indexing and Selection for Surveillance Video Summarization, IEEE 2019.

3. Khan Muhammad; Tanveer Hussain; Sung Wook Baik, Year: 2018, "Efficient CNN based summarization of surveillance videos for resource-constrained devices, Elsevier 2018.

4. Po Kong Lai; Marc Decombas; Kelvin Moutet, Year: 2016, "Video Summarization of Surveillance Cameras", IEEE 2016.

5. Sinnu Susan Thomas; Sumana Gupta; Venkatesh K. Subramanian, Year: 2017, "Smart Surveillance Based on Video Summarization", 2017 IEEE Region 10 Symposium (TENSYMP).

6. Wenzhong Wang; Qiaoqiao Zhang; Bin Luo; Jin Tang; Rui Ruan; Chenglong Li, Year: 2017, "Selecting Attentive Frames From Visually Coherrent Video Chunks For Surveillance Video Summarization", IEEE 2017.

7. ShuZhang; Yingying Zhu; Amit K. Roy-Chowdhury, Year: 2016, "Context-Aware Surveillance Video Summarization, IEEE Transactions On Image Processing, Vol: 25, no: 11.

8. Chen, Year: 2017, "Video to text summary: Joint video summarization and captioning with recurrent neural networks", In BMVC, pp. 1 – 10.

9. Mohamed Elfeki and Ali Borji, Year: 2019, "Video summarization via actionness ranking", Winter Applications in Computer Vision (WACV).

10. Chenyou Fan; Jangwon Lee; Mingze Xu; Krishna Kumar Singh; Yong Jae Lee; David J Crandall; Michael S Ryoo, Year: 2017, "Identifying first-person camera wearers in third-person videos", arXiv preprint arXiv:1704.06340.

11. Zhong Ji; Kailin Xiong; Yanwei Pang; Xuelong Li, Year: 2017, "Video summarization with attention-based encoder-decoder networks", arXiv preprint arXiv:1708.09545.

12. Zhong Ji; Yuanyuan Zhang; Yanwei Pang; Xuelong Li, Year: 2018, "Hypergraph dominant set based multi-video summarization. Signal Processing, Vol: 148, pp.114 – 123.

13. Zhong Ji; Yuanyuan Zhang; Yanwei Pang; Xuelong Li; Jing Pan, Year: 2019, "Multi-video summarization with query-dependent weighted archetypal analysis", Neurocomputing, Vol: 332, pp. 406 – 416.

14. Rameswar Panda; Niluthpol Chowdhury Mithun; Amit Roy-Chowdhury, Year: 2017, "Diversity-aware multi-video summarization", IEEE Transactions on Image Processing.

15. Behrooz Mahasseni; Michael Lam; Sinisa Todorovic, Year: 2017, "Unsupervised video summarization with adversarial LSTM networks", In Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp. 1 – 10.

16. Jingjing Meng, Suchen Wang, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan, Year: 2017, "Video summarization via multiview representative selection", In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1189 – 1198.

17. Rameswar Panda; Abir Dasy; Amit K Roy-Chowdhury, Year: 2016, "Video summarization in a multi-view camera network", In Pattern Recognition (ICPR), 2016 23rd International Conference on, pp. 2971 – 2976.

18. Rameswar Panda; Amit Roy Chowdhury, , Year: 2017, "Multi-view surveillance video summarization via joint embedding and sparse optimization", IEEE Transactions on Multimedia.

19. Mrigank Rochan; Yang Wang, Year: 2019, "Video summarization by learning from unpaired data", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7902 – 7911.

20. Yandong Li; Liqiang Wang; Tianbao Yang; Boqing Gong, Year: 2018, "How local is the local diversity? reinforcing sequential determinantal point processes with dynamic ground sets for supervised video summarization", In Proceedings of the European Conference on Computer Vision (ECCV), pp. 151–167.