VIDEO SUMMARIZATION USING MASK R-CNN

GIRISH PULINKALA[1] SAI SANKAR SRIRAM[2] SURYA WALUJKAR[3] PRANJALI THAKRE[4]

[1,2,3]Student, Dept. of Computer Engineering, SIESGST, Nerul, Maharashtra, India
[4]Asst Prof, Dept. of Computer Engineering, SIESGST, Nerul, Maharashtra, India

---***---

ABSTRACT:

The Data Revolution is going to play an essential part of our future. With a demand for improved data and a demand for better processing, data preprocessing plays a huge factor for analysis of data. Scouring through redundant data for useful and necessary information not only leads to wastage of time but also a huge waste of resources and personnel. It is essential that day by day we find methods and techniques that find us useful information from a huge pool of data so that analysis becomes easier and hassle free. Video Analysis is a tiresome task and searching for useful information in a huge video could take up a lot of time. Although many techniques exist to summarize videos and retain the most important information, we try to explore an algorithm Mask R-CNN to summarize videos. Through Mask R-CNN, we extract the essential keyframes from the existing video and try to combine them into a new, smaller video with essential and important information.

Key Words: Mask R-CNN, Keyframe extraction, RoI, RPN, FPN, ResNet, Bounding boxes, Video Summarization.

1. INTRODUCTION

Video summarization is the method of creating a summary of a video, in such a way that the video summary contains all the important information and instances from the video and it should be as compact and crisp as possible. The summarized video must show a good continuation between the frames, i.e. the summarized content must be a smooth addition of all frames extracted. Last but not the least, the video must not contain any redundancy and that depends upon the algorithm used and how well it has been trained to handle certain tricky scenarios.

Video summarization saves a ton of time and makes searching and analysis very easy. Companies, Organizations etc. can use this summarized video instead of going through the entire surveillance footage [7].

Video Summarization can use various methods and each method has its own set of advantages and disadvantages.

In our case, we plan to produce a video summary by using CNN, particularly Mask R-CNN which is an improvement over many algorithms including Fast R-CNN.

2. LITERATURE REVIEW

In the year 2015; Debi Prosad Dogra, Arif Ahmed and Harish Bhaskar proposed a smart video summarization technique that combined event(s)-of-interest occurring throughout all the frames that were a part of the video. Preceding segment average (PSA) and Cumulative moving average (CMA) were used as features that indicate sudden and gradual changes of the object with respect to the environment. [1]. The results of their experiments showed that their proposed method is invariant to the target(s) and were heavily dependent on the target(s) interaction with the environment and had reduced sensitivity towards global changes in scene and at the same time placed limited specification requirements on other local entities.[1]

In the same year i.e. 2015, Ross Girshick employed Fast R-CNN to efficiently summarize videos. Fast R-CNN uses previous frameworks and builds on them to efficiently classify object proposals using deep convolutional networks. Compared to previous work, Fast R-CNN makes use of several methods to improve training and testing speed while also increasing detection accuracy [2]. While this method is almost as efficient as any algorithm for video summarization, it poses a problem of stride quantization during ROI pooling. It causes misalignment and information loss. In January 2016, R. Girshick, J. Donahue, T. Darrell and J. Malik described the method to summarize videos by Region Based Convolutional Networks by accurately detecting objects and segmentation. Their approach combined two ideas: (1) Apply Convolutional Neural Networks to proposed regions in order to localize and segment objects and (2) when labeled training data are less, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, boosts performance significantly. Since its combined region proposals with onsite resulting model is called an R-CNN or Region-based Convolutional Network.[6]

In March 2017, Satpute, A. and Khandarkar, K. used the methodology of Key Frame extraction for effective video summarization. They listed various mechanisms for key frame extraction such as aggregation mechanism, Faber-Schauder wavelet, Haar Wavelet, Scale Invariant Feature Transform etc. All methods had their own sets of advantages, but these were outweighed by their disadvantages such as computational time, time complexity and their dependence on environment conditions.

In June 2017, Smt.M. Tirupathamma proposed frame difference as an option for key frame extraction and thus uses these key frames as a summary for the entire video. Key frames are extracted for video summarization using Frame Difference method. Key frames-based video summarization works on frames so initially a video frame sequence is divided into frames. The redundant content from the extracted video frames is discarded by calculating the frame difference between the adjacent frames. The frames whose difference is greater than certain threshold is considered as Key frames. The extracted Key frames are combined to form a summarized video [4].

In January 2018, came the improvised version of one

of the most effective algorithms that can be used for video summarization i.e. Mask R-CNN. Mask R-CNN, extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN [5]. Mask R-CNN outperforms all existing, single-model entries on every task [5].

In June 2019, a method involving keyframe extraction and video skimming was proposed as a technique to efficiently summarize a given video. Low level features for static keyframe were extracted using uniform sampling, image histograms and SIFT from Convolutional Neural Network (CNN). Clustering techniques such as K means and Gaussian Clustering were also employed. Video skimming was done in order to make the new video from the extracted keyframes smoother and fluid, thus making it easily comprehensible for humans. SIFT, VSUMM and CNN were able to summarize a video as close to a human summary i.e. more fluid and continuous with very less redundancy.

The mean scores for the videos suggested that CNN was performing better which was followed by VSUMM.SIFT was unable to outperform CNN and that is because CNN is more relevant and better for classification and categorization and has extremely good generalization abilities. It is currently more popular for image and video tasks [11] whereas SIFT can be used for real time scenarios.

In order to collect data for our project, we needed a data set. Although there are many sampling techniques that can be used, the one that suited us and our project implementation was convenience sampling.

3. SYSTEM ARCHITECTURE

As reviewed from the previous proposed system, the following drawbacks were:

1. Motion based algorithms would fail when pedestrians are still.

2. Fast R-CNN was not able to put a binary mask which differentiates whether the pixel belongs to the object.

3. ROI pooling slightly misaligns from the regions of the original image

4. Comparing video frames and extracting frames using frame difference takes up quite a lot of time.

Our proposed system uses Mask R-CNN, an uncomplicated, pliable, and general framework for object instance segmentation in deep neural networks. Mask R-CNN, is an extension of Faster R-CNN is made by branching and predicting an object mask in parallel manner with the existing branch for bounding box recognition. Mask R-CNN is easy to train and adds a small overhead to Fast R-CNN, which runs at 5 fps.

Mask R-CNN generates a binary mask which outputs whether the given pixel is part of the object. An input image and multiple regions of interest (RoIs) are input to a fully convolutional network. Each RoI is pooled into a fixed-size feature map and then mapped to a feature vector by fully connected layers (FCs). The network has two output vectors per RoI: SoftMax probabilities and per-class bounding-box regression offsets.

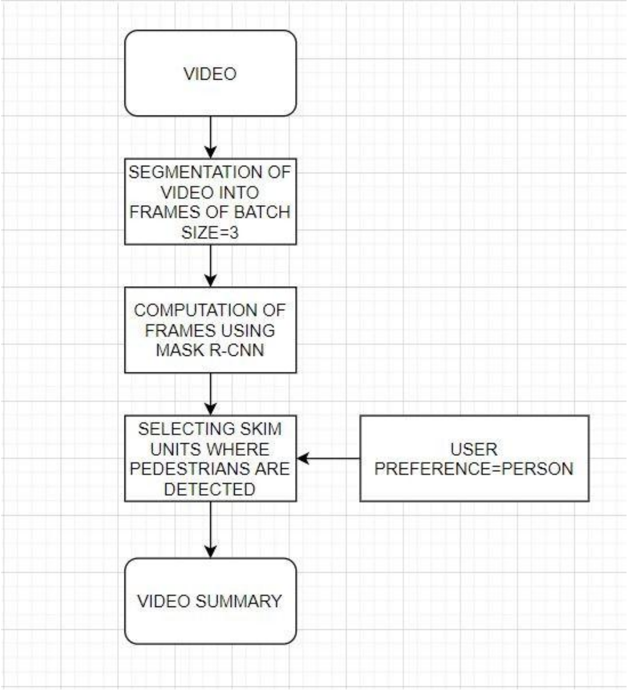
VIDEO
SEGMENTATION OF VIDEO INTO FRAMES OF BATCH SIZE=3
COMPUTATION OF FRAMES USING MASK R-CNN
SELECTING SKIM UNITS WHERE PEDESTRIANS ARE DETECTED
USER PREFERENCE=PERSON
VIDEO SUMMARY

Fig 3.1-Flowchart of Project

4. DETAILED WORKING

In this project, we are using keyframe extraction and video skimming for video summarization. For static keyframe extraction, we generate a binary mask which outputs whether the given pixel is part of the object in this we are considering the pedestrians. The input image and multiple regions of interest (RoIs) are then fed to the fully convolutional neural network (CNN) trained on MS COCO [Microsoft Common Objects in Context]. We apply video skimming by identifying the pedestrians on the cctv footage to make a summary video comprehensible for humans.

A) Keyframe Extraction:

1. Convenience Sampling:

Convenience sampling is a type of non-probabilistic method where the first available data is used as the primary data for key-frame extraction. Here, we use the pedestrians as the primary data. The idea is to select the key-frames wherever the pedestrians are present in the cctv footage by removing the redundant frames in which there is no extra information present. Convenience sampling is considered as the baseline feature of our project.

2. Mask R-CNN:

Mask R-CNN is a simple, flexible and general framework which is used in objects present in the image and predicts the class of the object from the mask. Using the predicted mask, we check for the pedestrian class to be present in the frame, if present then the frames are stored into the folder location defined or else if they are not present it discards the frames and starts the prediction for the next batch of frames. It not only provides the network to do the object detection but also pixel wise instance segmentation. Mask R-CNN has two stages:

I. Generating region proposals and classifying the proposals

II. Generating bounding boxes and object mask

In generating bounding boxes and masks, we are providing the Mask R-CNN model with pre-trained weights which were trained on MS-COCO. In our experiment the video is divided into frames of batch size three. These frames are then fed to the Mask R-CNN model which detects objects present in the image and predicts the class of the object from the mask. Using the predicted mask, we check for the pedestrian class to be present in the frame, if present then the frames are stored into the folder location defined or else if not present it discards the frames and then start the prediction for the next batch of frames.

B) Video summarization:

The approach is influenced by high level feature-based technique which uses the type and number of features for determining the representation of a frame and identifying the visually important contents. High level feature-based video summarization can include features such as object recognition, event detection in better understanding of the contents of the video. In our project the type of video skimming used is Generic Video Skimming System which proposes segmentation of the given video into smaller units known as skim units. Segmentation is initially performed by detecting the pedestrians in the frames of video. Segmentation helps in reducing the computation time as a representative frame is used instead of processing all the frames in the segments, in our case the batch size defined is three; which means one representative frame for three different frames.

After extracting the keyframes the fraction of the frames was used for reducing the computation time for video segmentation. The sequence of frames is strongly correlated, hence the difference from one frame to the next is expected to be very low when sampled at high frequencies. We used 30 frames per second as a sampling rate for our experiments and discarded the redundant frames. High level feature-based technique is time consuming and is computationally expensive.

5. ALGORITHM

Mask R-CNN is a simple, flexible and general framework which is used in objects present in the image and predicts the class of the object from the mask. It provides the network to do the object detection but also pixel wise instance segmentation.

5.1 Region Proposal Network:

Regional proposal network is a fully convolutional network which predicts the object scores and object bounds at the same time. RPN takes an image as an input and gives rectangular bounded boxes as the output each which has an object score, which is run on the fully convolutional network.

To generate the region proposal network, a small network is slid into a convolutional feature map which is the output of the last convolutional layer.[19] The small network takes an input of n x n window which is known as a sliding window.

Anchor:

Anchor is the center of a sliding window. The anchors are assigned labels based on features as:[19]

A. The anchors with highest Intersection-over-union overlap with a ground truth box.[19]

B. The anchors with Intersection-Over-Union Overlap higher than 0.7.

Anchor can be either Foreground or Background Class.

Loss Function:

The RPN is an algorithm which needs to be trained hence the loss function for an image is defined as: L({pi}, {ti}) = 1 Ncls X i Lcls (pi, p∗ i) +λ1 Nreg X i p ∗ i Lreg (ti, t∗i)

where;

i is the index of the anchor, Pi is predicted probability of the anchor i for being an object, p*i is the ground truth label,1 if the anchor is positive or 0 if the anchor is negative, ti is a vector representing the 4 parameterized coordinates of the predicted bounding box, and t ∗i is that of the ground-truth box associated with a positive anchor.

5.2 ROI Align:

The RPN returns the output proposals, all proposals are the offsets for each anchor, with the help of which we get the proposed bounding boxes. The given feature maps are n times smaller than the original image. It divides each coordinate(x) by n and by taking the float values: [x/n] it gives us the coordinates relative to feature map size. To extract the supposed object from the feature map we use the new coordinates for cropping. For getting the fixed size coordinates from the RoI pool the cropped part is divided into bins which is known as quantization. The RoI align divides the cropped part into a grid in which it selects four points through bilinear interpolation. The output of RoI align plays a key role in mask prediction.

5.3 Mask Representation:

A mask is a spatial layout of the object present in an image. Unlike the bounding boxes of the region of interest the it is collapsed into short vectors which uses fully-connected layers (*fc layers*). The m x m mask is predicted from the RoI regions using FPN. This allows each layer in the mask branch to maintain the explicit m × m object spatial layout without collapsing it into a vector representation that lacks spatial dimensions. The pixel to pixel behavior is determined by the RoI Align outputs which are the small feature maps.

5.4 Network Architecture (Resnet 101):

Deeper neural networks are difficult to train, hence to ease the training of networks a residual learning framework is used. Network Architecture is used to flow information from the layers in the model to next layers. ResNet was developed in the year 2015 which is short for Residual Network. ResNet is a classic neural network used as a backbone for computer vision tasks. ResNet uses skip connection to add the output from an earlier layer to a later layer. It helps in reducing the vanishing gradient problem by allowing this alternate shortcut path for gradient to flow through. ResNEt has a novel architecture with "skip connections" and features heavy batch normalization. ResNet can be 152 layers deep. In our project we have used the ResNet 101.ResNet 101 is a convolutional network which is 101 layers deep.

6. IMPLEMENTATION & RESULTS

In our Project we are applying Mask R-CNN on the videos from CCTV surveillance to summarize the video by removing redundant frames which provide no information other than the fixed background. The video summarization is done on the basis of the pedestrians or other moving objects where human presence is present. This technique of selecting the primary data or object and providing the prerequisite information is considered to be convenience sampling. At first, we are capturing the CCTV footage and dividing the video into frames of batch size of three, then we use the technique of keyframe extraction on basis of the sampling done on the video in our case it is pedestrians. Mask R-CNN is applied on the video frames to create a binary mask and to predict the class and label of the bounding boxes. If there are any pedestrians present in the video then the frame is saved into the folder or else if the frame does not contain any additional objects other than the background used then the frames are discarded. The whole video is then processed to find any objects until it returns no frames left. After extracting the keyframes, we used 30 frames per second as a sampling rate for our experiments and discarded the redundant frames since the keyframes are highly correlated to each other. The final video is presented which usually consist of the 10% time of the actual length of the video. The whole project was run on an Intel i7 CPU processor which captured the video divided into frames to feed it into the GPU processor. We used NVIDIA 940MX GPU processor which helped in processing the video and applying the Mask R-CNN on the video frames. It processes ~40 frames per minute which constitutes approximately 2 hours. The final video which was produced was 1920 x 1080 p.

The video of length 5 minutes and 16 seconds was used for the testing purpose of our project in which the masking and detection of every object in the frame was done.

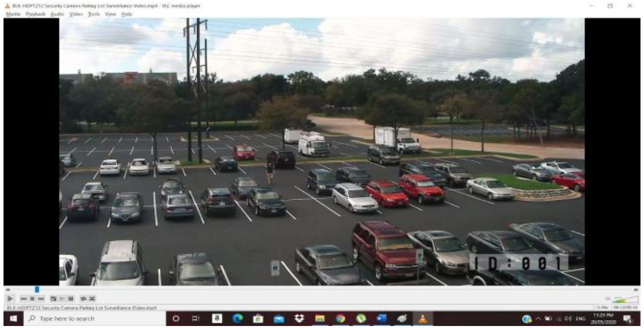

Fig 5.1-Length of Video

Masking of every object was done to detect the pedestrians from the video to remove redundant frames. Frame 120 was the first selected frame after removing the redundant frames.

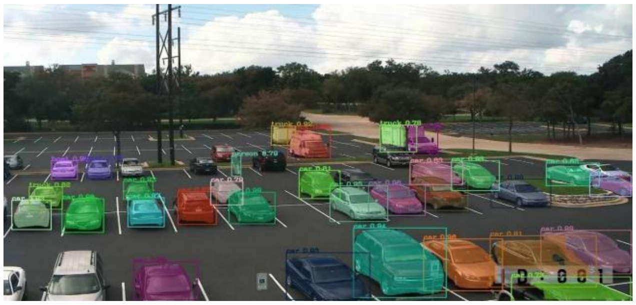

Fig 5.2-Frame 120 with mask

In our project for proper visual clarity we decided to remove the masking colors since and display the frame with only desired output, in our case it is the pedestrians.

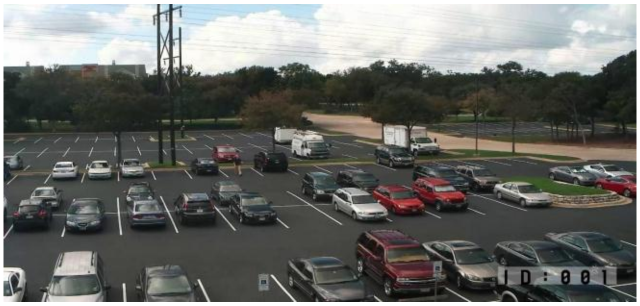

Fig 5.3-Frame 120 without mask

The output video length is 36 seconds which equals approximately to 8% of the total video length whose speed is 30fps.

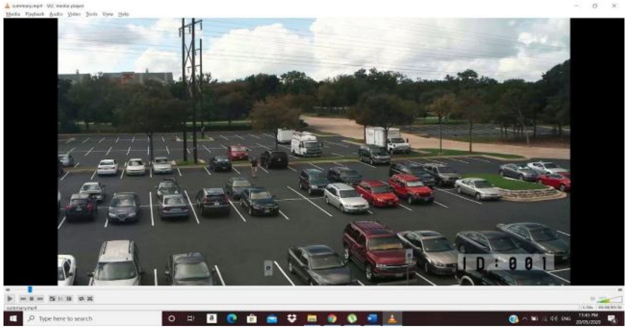

Fig 5.4-Length of Summarized Video

7. CONCLUSION

In this project we have introduced Mask-RCNN for video summarization. Our model is expected to deliver a high-performance video summarization system. Our proposed models are inspired by fully convolutional networks in semantic segmentation. We have adopted semantic segmentation networks for video summarization. Our model achieves a competitive performance in comparison with other supervised and unsupervised approaches that mainly use LSTMs. We believe that the Mask-RCNN model provides a promising and a better alternative to LSTM based approaches for video summarization. Using similar approaches, we can convert almost any semantic segmentation networks for video summarization. As future work, we plan to explore more verticals on recent semantic segmentation models and develop the counterpart models in video summarization.

8. ACKNOWLEDGMENT

We would like to express our sincere gratitude to Prof. Pranjali Thakre for her valuable guidance, support and help. We would like to thank her for constantly giving us feedback and assisting us in this project.

9. REFERENCES

[1] Dogra, Debi & Sk, Arif Ahmed Bhaskar, Harish. (2015). Smart Video Summarization using Mealy Machine based Trajectory Modelling. Multimedia Tools and Applications. 10.1007/s11042-015-2576-7.

[2] Ross Girshick (2015). Fast R-CNN. Microsoft Research.arXiv:1504.08083v2 [cs.CV]

[3] Satpute, A. and Khandarkar, K. (2017). Review on Video Summarization Techniques for Surveillance Video Using Keyframe Extraction. [online] http://ijaerd.com.

Available at: http://ijaerd.com/papers/finished_papers/Review%20on%20Video%20Summarization%20Techniques%20for%20Surveillance%20Video%20Using %20Keyframe%20Extraction-IJAERDV04I0327219.pd f [Accessed 30 Jul. 2019].

[4] Tirupathamma, S. (2017). Key frame-based video summarization using frame difference. [online] http://ijicse.in.Available at: http://ijicse.in/wp-content/uploads/2017/07/37.pdf [Accessed 17 Aug. 2018].

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick. (2018). Mask R-CNN. Facebook AI Research (FAIR). arXiv:1703.06870v3

[6] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pp. 142-158, 1 Jan. 2016. doi: 10.1109/TPAMI.2015.2437384

[7] UKEssays. November 2018. Video summarization techniques. [online].Available from: https://www.ukessays.com/essays/computer-science

/video-summarization-techniques.php?vref=1 [Accessed 3 November 2019].

[8] https://www.edureka.co/python

[9] https://github.com

[10] https://www.medium.com/@alittlepain833/simp le-understanding-of-mask-rcnn-134b5b330e95

[11] https://tams.informatik.uni-hamburg.de/lectures

/2015ws/seminar/ir/pdf/slides/JosipJosifovski-Object_Recognition_SIFT_vs\

_Convolutional _Neural_Networks.pdf

[12] Video Summarization using Keyframe Extraction and Video Skimming Jadon, Shruti Jasim, Mahmood. (2019).Video Summarization. 10.13140/RG.2.2.23087.38564/1.

[13] Taherdoost, Hamed. (2016). Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. International Journal of Academic Research in Management. 5. 18-27. 10.2139/ssrn.3205035.

[14] https://medium.com/@fractaldle/mask-r-cnn-un masked-c029aa2f1296

[15] https://docs.google.com/presentation/d/1NHdjqp5d2hbk9AIrN9Kp_AqCWupFCXN4cQQ3Whpc5wU/e dit#slide=id.g130b215d52_0_153

[16] https://www.analyticsvidhya.com/blog/2019/ 07/computer-vision-implementing-mask-r-cnn-image-segmentation/

[17] https://www.quora.com/What-is-the-difference-b etween-CNN-and-R-CNN

[18] https://www.researchgate.net/figure/Architectur e-of-ResNet-101-network-with-Dense-Upsampling-Co nvolution-DUC-layer_fig6_314115448

[19] Medium. 2020. *Region Proposal Network (RPN) — Backbone Of Faster R-CNN*. [online] Available at:

<https://medium.com/egen/region-proposal-network-rpn-backbone-of-faster-r-cnn-4a744a38d7f9> [Accessed 6 June 2020].