Video Object Segmentation and Feature Generation using Neural Network

Hari K.C.
PhD Scholar and Researcher
Kathmandu University
Dhulikhel, Nepal
harikc@wrc.edu.np

Sushil Shrestha
Assistant Professor
Kathmandu University
Dhulikhel, Nepal
sushil@ku.edu.np

Manish Pokharel
Professor
Kathmandu University
Dhulikhel, Nepal
manish@ku.edu.np

Abstract: Segmentation divides an image or a video into distinct regions that correspond to different objects or parts of objects. Object segmentation in the video is a challenging task as it requires segmenting objects into distinct regions over multiple frames. The aim of this work is the analysis and computation of features of objects extracted from video object segmentation. This paper uses Mask Regional Convolutional neural networks to extract objects and their features from video segments. The segmentation model is developed with Mask Regional-Convolution Neural Network that is trained on the available *Common Object in COntext (COCO)* public dataset. This dataset has 80 different categories of object. For testing purposes, YouTube video and local video datasets are used. The experiment is performed on an outdoor video dataset. The result shows that objects in the frames of the video are extracted and features of the objects are analyzed with an average accuracy of 85.53 percent. The significance of this research is to extract the features of a segmented object such as *shape*, *size*, *texture*, and *color*. These features are used for understanding the segmented objects in video and can be further used for video retrieval, browsing and summarization. Manufacturing companies, security organizations, media companies, government agencies and healthcare centers will be benefitted from the research.

Keywords: Video segmentation, Object analysis, Object recognition, Mask R-CNN, feature generation

I. INTRODUCTION

Video segmentation is an emerging area of computer vision. Video object segmentation is the process of separating the foreground objects in a video sequence from the background [1]. It is a crucial step in many computer vision applications, including object tracking, action recognition, and scene understanding. The activity of video object segmentation is challenging due to the presence of motion, occlusions, and changing illumination conditions. Segmentation is a video content analysis task that can be further used in different applications such as video compression, video surveillance and video summarization. It can be divided into instance and semantic segmentation for object recognition and analysis. Instance segmentation is a task where the goal is to identify and segment individual objects within a video frame, as well as assign a unique label to each object. Instance segmentation combines object detection and semantic segmentation, where each instance of an object is treated as a unique object and segmented accordingly. On the other hand, semantic segmentation assigns a semantic label to each pixel in an image or video frame [2]. Unlike instance segmentation, semantic segmentation does not differentiate between instances of the same class. Instead, the focus is on classifying the pixels in the image into semantic categories, such as people, cars, buildings, etc. The aim of semantic segmentation is to obtain dense labeling of the image, where each pixel is classified into its corresponding semantic category. Object analysis and feature extraction refer to the process of automatically extracting meaningful information about objects in an image or video [3]. This information can be in the form of features such as shape, size, texture, color, or other attributes that can be used for further analysis and processing. Object analysis is an important step in many computer vision tasks, such as object detection and segmentation, as it provides a compact representation of the objects in an image that can be used for various purposes such as recognition, tracking, and classification. Feature extraction is the process of transforming raw data into a compact and meaningful representation that can be used for further analysis in video compression, video surveillance and video summarization. It is used to extract video frame features such as edge detectors, texture descriptors, or color histograms, which can be used as input to a neural network for object recognition, and analysis. Traditionally, video object segmentation for object recognition and analysis was performed using hand-crafted features and models, such as optical flow and Gaussian mixture models. With the advancement of deep learning, more sophisticated models, such as Mask R-CNN, have been implemented. The Mask R-CNN architecture, which is based on the ResNet architecture, has proven to be effective in various computer vision tasks, including instance segmentation. This paper uses Mask R-CNN for video object segmentation and evaluate its performance on the COCO dataset. The COCO dataset is a large-scale image recognition, segmentation, and captioning dataset [4]. It contains more than 330,000 images and over 2.5 million object instances. The dataset includes 80 object categories such as Person, Bicycle, Car, Motorcycle, Airplane, Bus, Train, Truck, Boat, Traffic light, Fire, Stop sign, Parking meter, Bench, Bird, Cat, Dog, Horse, Sheep, Cow, Elephant, Bear, Zebra, Giraffe, Backpack, Umbrella, handbag, tie, suitcase, Frisbee, Skis, Snowboard, ball, kite, bat, glove, skateboard, surfboard, Tennis racket, bottle, wine glass, Cup, Fork, Knife, Spoon, Bowl, Banana, Apple, sandwich, orange, Broccoli, Carrot, Hotdog, Pizza, Donut, Cake, chair, couch, plant, bed, dining table, toilet, TV, laptop, mouse, Remote, keyboard, cellphone, Microwave, Oven, Toaster, Sink, Refrigerator, Book, Clock, Vase, Scissors, Teddy bear, Hair dryer, Toothbrush.

Despite the significant progress that has been made in the field of video object segmentation, the task of object detection feature computation is a challenging problem. Several issues such as objects can appear in different scales and orientations make it difficult to detect consistently across different segments of the videos and continue to be a source of difficulty. Similarly, Objects appear in dynamic outdoor environments making it difficult to track them. Hence, to address and overcome these problems and challenges, this paper provides the video segmentation model for digital video using ResNet architecture [5]. The objectives of this research are as follows:

a) To implement and evaluate the performance of mask R-CNN for video object segmentation.

b) To extract the object features from video segments.

II. RELATED WORK

The history of video object segmentation dates back to the early days of computer vision, where the goal was to segment objects into a single image. However, with the advent of video technology, the need for segmenting objects in videos became increasingly important. This led to the development of video object segmentation methods [6][7] that could handle the temporal information in videos and segment objects over time. One of the earliest approaches to video object segmentation was based on optical flow, which uses the motion information between consecutive frames to segment objects. However, this method was limited by the accuracy of optical flow algorithms and the difficulty of handling object occlusions and deformations. In recent years, deep learning-based methods, such as instance segmentation models like Mask R-CNN, have become popular for video object segmentation. These methods use convolutional neural networks to learn the complex relationships between pixels and objects in videos, providing more accurate and robust results compared to traditional optical flow-based methods.

In 2017, one of the earliest studies proposed a video segmentation method based on pixel-level matching using CNN [8]. The proposed network shows objects using features from separate layers to leverage both spatial detail and category-level semantic information. Experiments on large datasets demonstrate the efficacy model compared to related methods in terms of accuracy, speed, and stability. In 2018, the authors [1] explained the segmentation of actors and their actions of video content at pixel-level. They infer segmentation from a natural language input sentence. The proposed architecture was an encoder-decoder with fully convolutional network. Datasets with greater than 7,500 natural language descriptions were taken. Experiments show that the feasibility and robustness, as well as the model's ability to adapt to the task of semantic segmentation of actor-action pairs, results in an average accuracy improvement of 8.8% and an accuracy improvement of 7.2% for the state-of-the-art technique. In 2018, the researcher developed a video segmentation framework that includes two novel components, a feature propagation module for reducing the cost of per-frame computation and an adaptive scheduler that allocate computation based on accuracy prediction. The experiment was performed on Cityscapes and CamVid datasets. The results on both Cityscapes and CamVid showed that their method yielded a substantially better tradeoff between accuracy in latency [9]. In 2019, A research paper [10] considers an orthogonal approach that processes each frame independently without considering temporal information. The authors presented a one-shot semantic segmentation of video objects based on a fully convoluted neural network architecture that can continuously transfer general semantic information. They used the DAVIS data set containing 50 annotated full HD video sequences (30 for training set and 20 for validation set). The result adds robustness to the appearance changes of the object and helps in keeping the quality throughout a longer period of the video.

Similarly, in 2020, another study by authors [4] at Boston University designed distributed networks for fast video segmentation. The features extracted from high-level layers of a deep Convolution neural network can be determined by composing features extracted from several shallow networks. They perform the experiments using datasets such as Cityscapes, Camvid for street views, and NYUDv2 for indoor scenes. The result retains high accuracy while significantly improving the latency of processing video frames. Again, In 2020, a novel unsupervised image segmentation was proposed by authors that consist of normalization and an argmax function for differentiable clustering [11]. Experimental results on segmentation benchmark data set PASCAL VOC 2012 and BSD500 showed the effectiveness of segmentation in unsupervised learning. The proposed technique transcends traditional unsupervised image segmentation methods such as k-means clustering and graph-based segmentation methods that validated the importance of feature learning. However, in 2021, researchers [2] proposed a temporal memory attention network for adaptively integrating long-term temporal relationships across video sequences based on self-attention mechanisms without using full optical flow prediction. Experiments were run on two benchmark datasets, Cityscapes and CamVid. Cityscapes contains 5000 high qualities, detailed annotated images. CamVid[12] contains 4 videos with 11 category labels. Annotated frames are grouped into snippets of 467, 100, and 233 for training, validation, and testing, respectively. The proposed method achieved state-of-the-art performance on Cityscapes and CamVid dataset without complicated testing augmented skills. The objective of this literature review is to survey the recent advancements and current state of art in the field of video object segmentation using neural networks and to provide an overview of the key techniques and approaches used. However, there is still much room for improvement, and further research is needed to develop more accurate and robust models for video object segmentation and feature generation.

Overall, the use of deep learning techniques in video segmentation holds great promise for improving the understanding of player performance and fan engagement.

III. Framework, Methods and Tools

A. *Research Framework*

The overall framework for this research is depicted in Figure 1. It consists of video to frame conversion, video preprocessing and video object segmentation and detection using mask R-CNN.

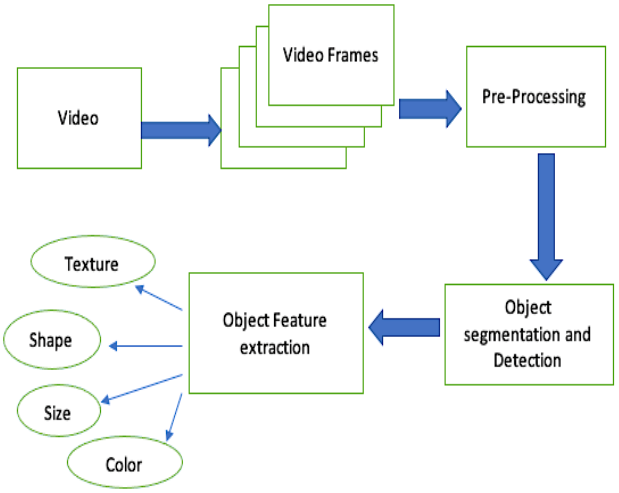

Fig. 1: Conceptual Framework

B. *Methods and Tools*

a) Video to Frame Conversion

Video contains metadata or additional information other than the image itself such as created date, device information and other technical information. To get a frame from a particular point of a video, an important piece of information, number of frames per second or fps is used. Video taken for an experiment is 24 fps. These frames are ordered and can be found by the frame number. Then using the number of frames per second, and the time in the video, the frame number can be easily calculated and extracted from the video. Finally, those frames are saved for further preprocessing, object segmentation, and feature extraction.

b) Video Frame Preprocessing

It consists of cropping and resizing video frames. Cropping and resizing a frame in video object segmentation involves selecting a portion of the frame and changing its size. The cropping process involves selecting a portion of the frame by defining a rectangular region of interest (ROI) that contains the object of interest. ROI (Region of Interest) is a concept used to select a subset of feature map regions that correspond to objects in an image. The ROI is defined as a rectangular bounding box that encloses each object of interest in the image. The resizing process involves changing the size of the frame by specifying the desired width and height.

ROI = image [y:y+h, x:x+w] -------(1)

Here, "x" and "y" typically refer to the horizontal and vertical coordinates of a pixel or a region in the image, respectively. "h" and "w" are the height and width of a region, such as a bounding box, in the image.

Resized image = image (x1, y1) --------(2)

Here x1, y1 are new coordinates of pixel or region in image.

c) Object Segmentation and Detection

Regional-CNN is based on the Faster R-CNN architecture and uses a multi-stage process to perform object segmentation and detection.

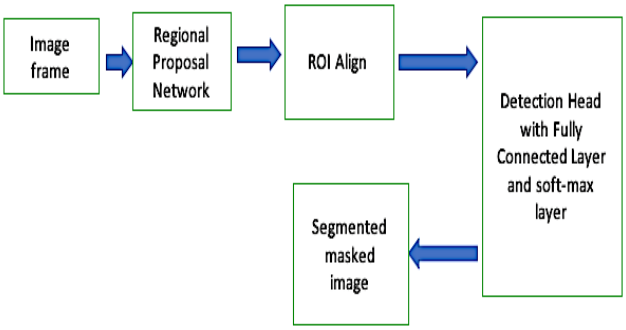

Fig. 2: Mask R-CNN

The main components of the architecture are:

a) Region Proposal Network (RPN): The Region Proposal Network (RPN) in Mask R-CNN [13] is a network that is used to predict object proposals. The RPN uses a combination of convolutional and fully connected layers to extract features from an input image. Then, it performs a sliding window operation to generate candidate object proposals. The RPN uses a 3x3 convolutional layer, followed by a ReLU activation, to generate the proposals for the objects. The proposals are generated by applying anchor boxes of different shapes and scales on the feature maps. The RPN can be represented mathematically as: **Binary classification score:** Let p_i be the predicted probability that the i-th anchor box contains an object, given by a 2-class softmax layer.

p_i = softmax(w_i * f + b_i) -------- (3)

where w_i and b_i are the weights and biases of the softmax layer, and f is the feature map output by the RPN.

Bounding box regression:

Let t_i be the target values for the regression coefficients of the i-th anchor box.Let y_i be the predicted regression coefficients for the i-th anchor box.

y_i = w * f + b -------- (4)

where w and b are the weights and biases of the regression layer, and f is the feature map output by the RPN.

The binary classification scores and the bounding box regression coefficients are used to generate the final object proposals. The proposals with the highest classification scores are selected, and the bounding box

regression coefficients are applied to adjust the anchor box locations to fit the objects more precisely.

b) RoIAlign: The RoIAlign operation in the Mask R-CNN architecture involves pooling and aligning the features from the region of interest (RoI) in the convolutional feature map. Given a set of regions of interest, the RoIAlign operation pools the features within these regions and aligns them to a fixed size, regardless of the scale of the original feature map. The pooling operation is performed with a bilinear interpolation operation. It uses a weighted average of the 4 closest pixels to estimate a new pixel value. The equation is given by:

output_value = (1-dx) (1-dy) * value_at_xy1 + (1-dx) * dy * value_at_xy2 + dx * (1-dy) * value_at_xy3 + dx * dy * value_at_xy4 ------- (5)

Where dx and dy are the fractions of x and y between the original and desired pixel locations, value_at_xy1, value_at_xy2, value_at_xy3, and value_at_xy4 are the values of the four closest pixels to the desired pixel location.

c) Detection Head: This processes the features from the RoIAlign and predicts the class labels, bounding box coordinates, and instance masks for the objects in the image. The detection head may use a fully connected layer to predict the class labels, followed by a SoftMax activation to produce the class probabilities. The softmax activation function is given by:

softmax(x_i) = e^(x_i) / Σ e^(x_j) ------ (6)

where x_i is the input value for the i-th class, and e^(x_i) is the exponential of x_i. The denominator is the sum of the exponentials of all classes, which ensures that the outputs of the function are normalized probabilities, i.e., the sum of all class probabilities is equal to 1.

The detection head in Mask R-CNN uses a fully connected layer to predict the class and location of an object. The output of the detection head is a tensor with shape **(N, (K+1) * 4),** where N is the number of predicted objects and K is the number of classes. Each row in the tensor represents a predicted object, with the first 4 values representing the bounding box location (x1, y1, x2, y2), and the rest of the values representing the confidence scores for each class.

IV. Data Analysis and Results

A. *Data*

For training the model, COCO public dataset is used and for testing purposes, video datasets are collected from YouTube and saved in a local database. Videos of the duration of up to 3 minutes are used for the experiment due to computing resources constraint. Mainly, outdoor environment videos are taken since these videos contain dynamic objects.

B. *Data Analysis Tools and Techniques*

For the analysis purpose, python programming language with computer vision and image processing functions are used. The experiment is conducted with videos from YouTube in Google Colab. Videos having 24 fps are used for the experiment. For evaluation, accuracy is computed.

C. *Analysis Results*

The outputs of the experiment are as follows:

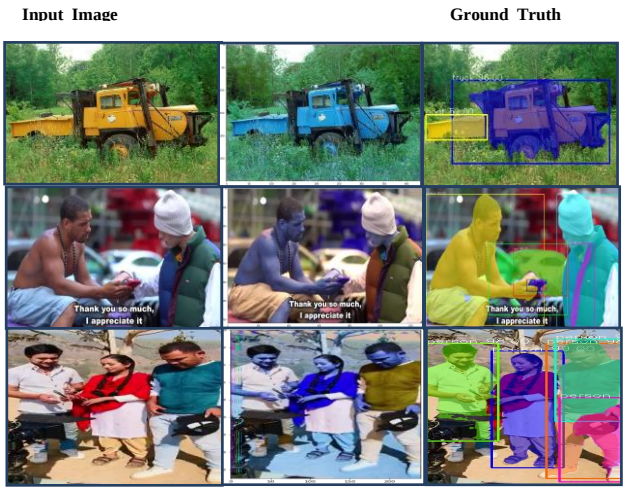

Fig. 3: Input, Ground Truth and Output Samples

Fig. 4: Segmented frame sample from sample video

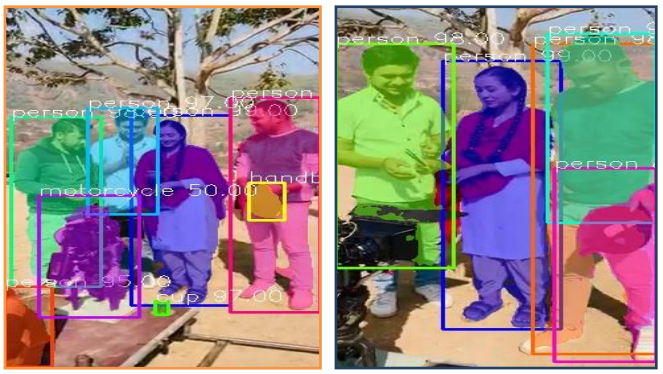

Fig. 5: Segmented sample from another video source

Feature extraction from detected object after object segmentation in given sample frame

a) Object 1(Person1): Accuracy = 99.00%
size: Area = 1431.5 pixels, Perimeter = 184.75 pixels
texture feature:
Contrast: 359.9369677711071
Homogeneity: 0.5919791580035698
Correlation: 0.9506749757498868
shape: Polygon

Color: RGB representation of a color with values [243.4, 212.22, 194.1]

b) Object 2(Person 2): Accuracy = 99.00%
size: Area = 268.4 pixels, Perimeter = 1394.5 pixels
Texture feature:
Contrast: 1120.939504067959
Homogeneity: 0.45786440256096855
Correlation: 0.8841929130288804
shape: Polygon
Color: RGB representation of a color with values [5.7, 244.65,2.99]

c) Object3
For object 3, segmentation model wrongly segments and predict/recognize the object in this frame. In reality, this object is the front layer of the jacket, not the tie**.** This is the exceptional case by the model due to the variation in object appearance and visual similarity. In other video frames, the model is working correctly.

d) Object 4 (Car 1):
Accuracy = 96.00%
Size: Area = 93.00 pixels, Perimeter = 852.20 pixels
Texture feature:
Contrast: 1188.226777962412
Homogeneity: 0.48698709310474286
Correlation: 0.9436629445492359
shape: Polygon
Color: RGB representation of color with values [(234.22, 245.44, 255.00)]

e) Object 5 (Car 2)
Accuracy = 62.00%
Size: Area = 31.5 pixels, Perimeter = 392.43 pixels
Texture feature:
Contrast: 797.5030222615362
Homogeneity: 0.42940986477801363
Correlation: 0.8120896837744835
shape: Polygon
Color: RGB representation of color with values [(234.22, 0, 2.55)]

f) Object 6 (cellphone)
Accuracy = 91.00%
Size: Area = 1.00 pixels, Perimeter = 4.82 pixels
Texture feature:
Contrast: 46.888888888888886
Homogeneity: 0.3851424955280332
Correlation: 0.7791505536306209
shape: Polygon
Color: RGB representation of color with values [(200.33, 1.44, 2.11)]

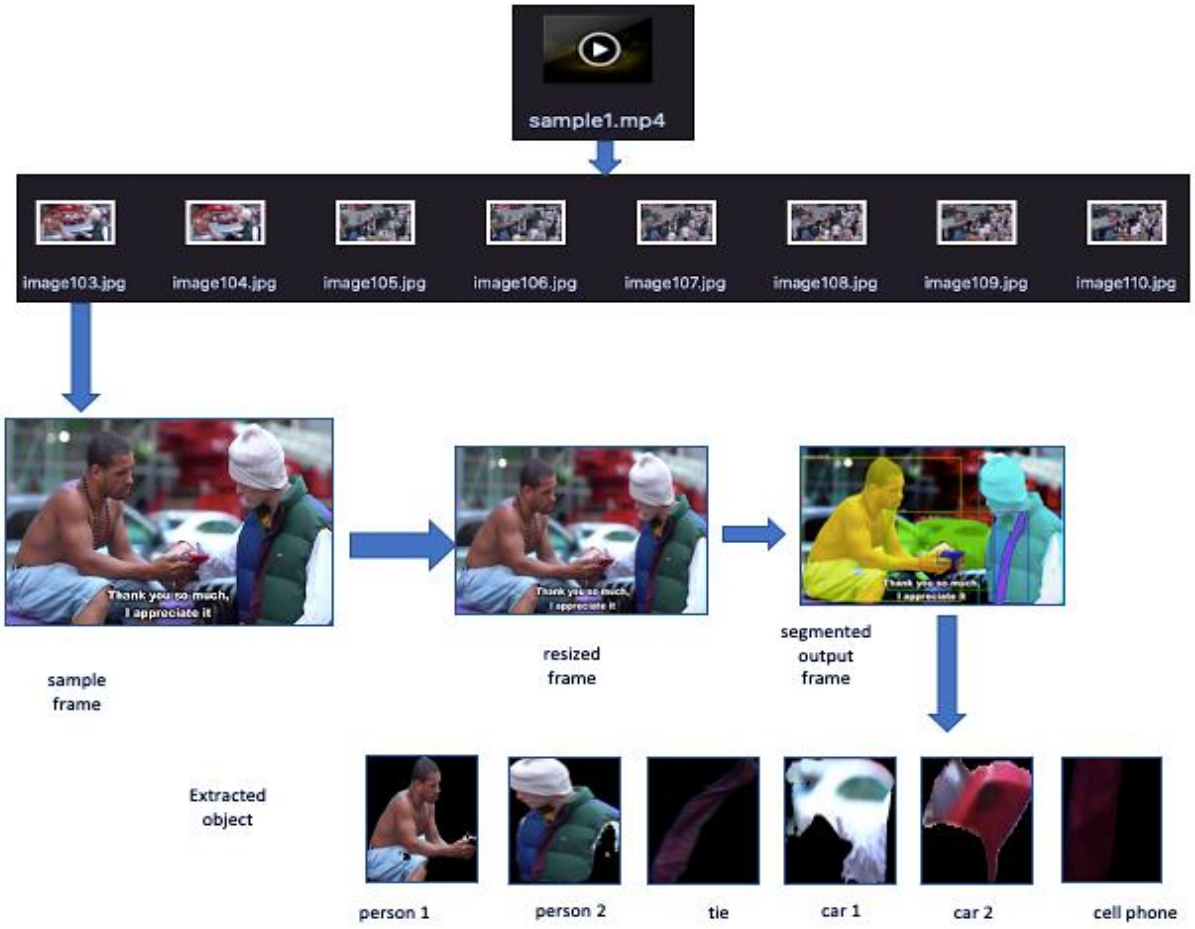

Fig. 6: Process and Result

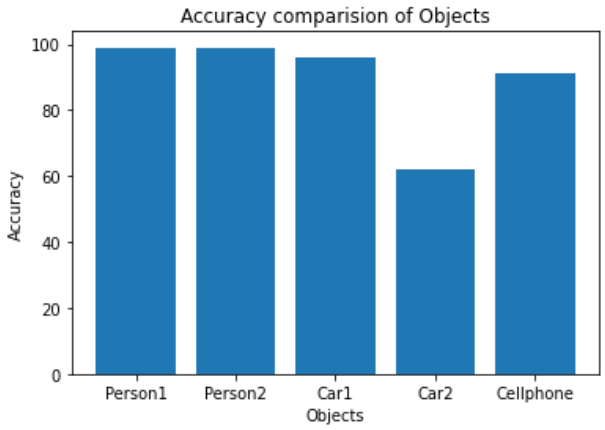

Fig. 7: Accuracy comparison of object in sample video

For evaluation, accuracy of the model has been obtained. It is computed by comparing output to ground truth segmentation pixel by pixel. The ground truth contains the objects of interest of the images. Accuracy is calculated as follows:

Accuracy(A) =[(TP + TN) / (TP + TN +FP + FN)]—(7)

The ratio of the total number of correct predictions to total number of predictions made by model.

Here, TP = True Positive, no. of instances where positive class is predicted correctly, TN = True Negative, no. of instances where negative class is predicted correctly = False Positive, no. of instances where positive class is predicted incorrectly, TN = True Negative, no. of instances where negative class is predicted incorrectly.

The average accuracy of the model is 85.53 %.

V. Conclusion and Future Works

The video object segmentation is performed using mask R-CNN. Segmentation is very precise since the objects are efficiently segmented in the video frame with high accuracy. The objects in the segmented frame are accurately recognized, however the model failed to recognize one of the objects in one experimented frame in the sample video. But, In the same video, segmentation model segment and predict accurately for other sample frames. The average accuracy obtained by the model is 85.53 %. The segmentation model has been tested for other video sources as well. The segmentation model works well with other sources as well.

The limitation of this segmentation model is that it doesn't work in visual similarity condition. Also, there are limited classes in the dataset. In the future, the recommendation will be, the model dataset can be customized with more local dataset including many data classes to train the model and make it more precise. Another recommendation will be optimization of parameters of the model. Similarly, the features obtained from this segmentation model can be used as input features to the deep learning model for the task such as video compression, video surveillance and video summarization.

References

[1] K. Gavrilyuk, A. Ghodrati, Z. Li, and C. G. Snoek, "Actor and action video segmentation from a sentence," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[2] H. Wang, W. Wang, and J. Liu, "Temporal memory attention for video semantic segmentation," 2021 IEEE International Conference on Image Processing (ICIP), 2021.

[3] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, and D. Puig, "Feature extraction and selection for emotion recognition from electrodermal activity," IEEE Transactions on Affective Computing, vol. 12, no. 4, pp. 857–869, 2021.

[4] T. Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, et al., "Microsoft COCO: Common Objects in Context," CoRR, vol. abs/1405.0312, 2014.

[5] P. Amatye, H.KC. "Performance Analysis and Classification of Rice Plant Disease using Multi- class Support Vector Machine and Transfer Learning Models", 2021 Proceedings of 10th IOE Graduate Conference,2021.

[6] P. Tokmakov, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[7] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos, "Efficient Video Object Segmentation via network modulation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[8] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos, "Efficient Video Object Segmentation via network modulation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[9] Y. Li, J. Shi, and D. Lin, "Low-latency video semantic segmentation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[10] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixe, D. Cremers, and L. Van Gool, "Video object segmentation without temporal information," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 6, pp. 1515–1530, 2019.

[11] W. Kim, A. Kanezaki, and M. Tanaka, "Unsupervised learning of image segmentation based on differentiable feature clustering," IEEE Transactions on Image Processing, vol. 29, pp. 8055–8068, 2020.

[12] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla, "Segmentation and recognition using structure from motion point clouds," in ECCV, 2008, pp. 44–57.

[13] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961-2969.

[14] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi, "Temporally distributed networks for Fast Video Semantic Segmentation," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in CVPR, 2016, pp. 3213–3223.

[16] J. S. Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, and I. S. Kweon, "Pixel-level matching for video object segmentation using Convolutional Neural Networks," 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[17] Y.-T. Hu, J.-B. Huang, and A. G. Schwing, "VideoMatch: Matching Based Video Object Segmentation," Computer Vision – ECCV 2018, pp. 56–73, 2018.

[18] W. Wang, J. Shen, F. Porikli, and R. Yang, "Semi supervised Video Object Segmentation with Super-Trajectories," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 4, pp. 985–998, 2019.

[19] Y. Li, J. Shi, Dahua Lin. "Low-Latency Video Semantic Segmentation", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[20] K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixe, D. Cremers, L. Van Gool. "Video Object Segmentation without Temporal Information", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.