INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 5 - V7I5-1381)
Available online at: https://www.ijariit.com

Video Shot Segmentation: Hybrid Approach using YOLOv4 and Deep Sort Algorithm

Shakthi T.
shakthi011001@gmail.com
Vellore Institute of Technology, Vellore, Tamil Nadu

ABSTRACT

A shot is a sequence frames in an edited video taken by a single camera. Shot Segmentation is the process of splitting video and finding the boundaries of video data. In this paper, we study about method in content-based video retrieval which uses object detection and tracking for video segmentation. Data collected via segmenting can be categorized in hierarchy manner as scene layer, camera shot layer and the frame in their accordance. The data collected is used for segmentation. YOLOv4 is used to enhance the accuracy and the process of tracking and detection much faster.

Keywords— *Video shot segmentation, Video object detection, Boundary detection, Object tracking, YOLOv4, Deepsort*

1. INTRODUCTION

Many traditional computer vision algorithms for segmentation based on background subtraction techniques, which are affected a lot by light-switch and targets' movements. On the other hand, most of the modern deep learning-based frameworks run at a deficient speed that cannot support real-time usage, although they have better robustness against scene changes. The proposed model, SEG-YOLO instance segmentation, is an extension of YOLO (You Only Look Once) version 4, which is one of the state-of-the-art object detection models. The extension part is FCN (Fully Convolution Network), which is used for semantic segmentation. SEG-YOLO aims to overcome both the speed and accuracy problems, while its usage can also be generalized to much extent.

With the development of multimedia technology, video resources the demand for the daily creation of digital videos is of an outsized range. Therefore, as the demand grew exponentially alongside the ascension of the net, the sector of video categorization, classification, retrieval, and segmentation turned into a vigorous space of analysis. Manual edits provided out to extremely time intensive and consuming task in hands, Automation's significance started to kick in where applications which have made the requirement to perform functionalities such as modeling, indexing, retrieving, browsing, segmenting and many such others in the massive multimedia system knowledge.

Many modern automation technologies require multi-sensor inputs for correct operations. Computer vision is one of the most important parts, and it has been applied in several areas, including autonomous driving, surveillance, and watching experience enhancement on broadcast TV. The system focuses on still camera vision for watching, experience enhancement, and precise segmentation on videos.

YOLO is derived from a single Convolutional Neural Network (CNN). The CNN divides a particular image into boundaries and regions and then it predicts the boundary boxes and probabilities for each region. It simultaneously also predicts multiple bounding boxes and probabilities for those classes. It also sees the complete image during the training stage and tests time, so it encodes the relevant information about classes as well as how they look implicitly.

The direct application of the work comes in terms with analyzing the frames of a video in the vision community that is related to the object that needed to be detected. Video shot segmentation is a merger of three image detection tasks, classification, localization, and segmentation. In this concept, the whole system can be classified by two terms: (a) object detection, which contains classification and object localization; (b) semantic segmentation that generates a mask of the object. Each of the techniques is previously well studied and implemented as well.

In summary, the contribution involved while making the proposed system are:

- Construction of a novel CNN architecture for video shot segmentation. Our model runs in real-time at 2% more accurate relative to the traditional algorithms.
- Implementation of a loss function improving the process of localization of entities into separate segments
- Demonstration of how our hybrid generalizes to tasks with various other datasets and proves to be fast and accurate in the tracking and detecting process.

© 2021, www.IJARIIT.com All Rights Reserved

The rest of the paper is being categorized as listed. First, we discuss the related work, followed by the architecture and implementation details. Then we present the details of the creation of the model and the experimental setup and discuss the results. Finally, we conclude with the main findings in the conclusion section of the paper.

1.1 Related Works

In work [1] they have used variational mapping, effective learning of class labelled embedding vectors by mapping from semantic space to visual space to do zero shot semantic segmentation, which on training has never seen target class.

In work [2] they have shown the segmentation methods of shot boundary using conventional ML and DL techniques. Speeded up Robust Features based video shot segmentation algorithm is found on [3] that computes SURF features of video frames obtaining the boundary of shots which calculates feature machine rate. They created bag of visual world in [4] visual scenes of video segmentation which are divided into shots represented by set of key frames. Unsupervised shot detection method has been used in [5] to classify news and commercial video shots using a smaller number of parameters by shot segmentation and shot classification.

In work [6] we can see the video object segmentation (VOS) which makes unlabeled videos address object learning pattern. They used supervised learning framework to capture multiple granularities of intrinsic properties of VOS. A statistical model can describe these kinds of characteristics, for example, a Gaussian Mixture Model (GMM). The GMM combines multiple weighed Gaussian models and imitates/replicates the background. Once the parameters of the model is calculated, the model can present the background. The foreground object then can be detected by distinguishing pixels that do not belong to the background model. This process is known to be background subtraction. Zivkovic [8] developed a GMM based background subtraction algorithm. However, the real world's scene is volatile. Many factors could change, including light-switch, shadow, leaf-swing, raindrop, noise, etc. These changes could result in an unfavored segmentation. MOG2 [9] is a modified GMM based background subtraction algorithm, it mainly improved in (a) Add shadow detection function; (b) Use adaptive GMM against environment changes. Yet, it is still not good enough in our use case. Mask R-CNN [10] is an instance segmentation model by combining object detection and semantic segmentation. It uses object detection to distinguish different objects, and then for each different object, it performs semantic segmentation. Mask R-CNN has extraordinary performance, and it is always referred to as baseline in recent papers.[11] Involved in the process of adjusting of the connections in the network so as to minimize a measure of the difference between the output vector and the desired output vector

2. ARCHITECTURE

SEG-YOLO instance segmentation model can be separated into 3 parts. The first part of the process is YOLOv4, which proposes the region of interest (ROI) and retrogresses for the classes and confidence scores. In the proposed system, a hybrid YOLOv4 is implemented, which also outputs the feature maps from every layer of the left blocks. The second part of the process is Region of Interest (ROI) Align that takes a different YOLO bounding box output and the feature map as input and would generate the fixed-size ROI feature map. The last part is involving the Fully Convolution Network (FCN), which indulges the process of transforming the ROI inputs to semantic mask outputs.

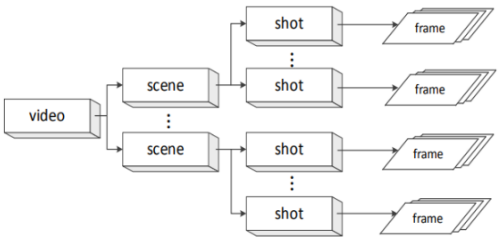

Figure-1 Video shot segmentation structure

So a classifier is build based on the dataset provided which is further trained it achieves a reasonably good accuracy. Then we grab this network and connect it with the final classification layer leaving behind a dense layer that would produce a single feature vector, waiting to be classified. The feature vector that is mentioned above is also known as the appearance descriptor. Thus we make the Hybrid of YOLOv4 and Deep sort which will further help in enhancing the Object detection process of Segmentation.

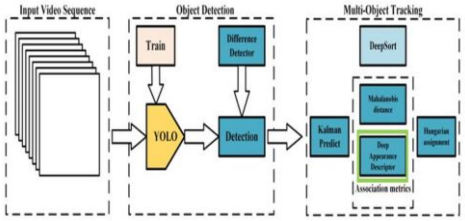

Figure-2 Deepsort Architecture

You Only Look Once(YOLO) is an end-to-end network for object detection. YOLO splits the image to SxS block, and then each block is responsible for detecting those targets whose center points fall within the grid. After detection, Non-Maximum Suppression is used for eliminating the duplicated bounding boxes. The 4th generation of YOLO, YOLOv4 has integrated many cutting-edge technology, including residual block based backbone, feature pyramid network like network head for multi-scale prediction, batch normalization, anchor boxes prediction, etc.

The cross-entropy between training data and the model's detection is usually the cost function of the DNN in classification problems. Training a DNN means adjusting the neuron's connection weights to minimize the cost function.

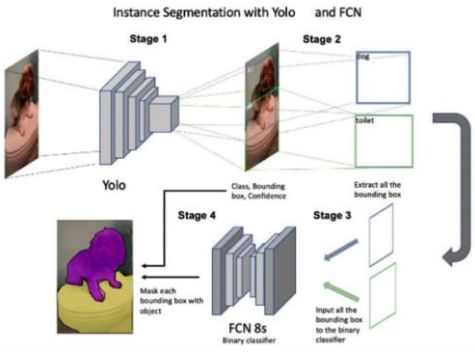

Figure-3 Segmentation Architecture

3. METHODOLGIES/TECHNIQUES

We have used 1000 images of bus, car, truck, person and bicycle each for the training dataset. And used 200 images of the same classes each for testing the trained dataset.

© 2021, www.IJARIIT.com All Rights Reserved

In this stage, the prediction variable was sliced to include only the bounding box coordinates and changed back to integer values in order to get the correct coordinates. The bounding box image was an array due to the slicing from the original image. Transforming to tensor was carried out, which fulfils the Pytorch requirement.

FCN-8 had been already trained on a dataset before to produce very good results. However, in our case we are interested in using FCN-8 as a binary classifier (only generating a mask for the background and object). For this to work the last six convolutional were modified to make predictions for two classes only. As results new weights are generated using the He method of initialization. Furthermore, the data used to train the model was slightly modified to work for binary classification. Since we are no longer interested in the model's ability in distinguishing between different classes in the image, the ground truths are modified to produce a single mask for all the classes in the image. All images in model were resized to (408 X 408). Also, the pre-processing of the images followed that of the research behind FCN-8 in which the per channel mean is subtracted from the image.
Lastly, we combine all of the detection and FCN to form the instance segmentation. Mask FCN8s outputted a 2 layers tensor with shape equal to the inputting dimension. In here, we detached the tensors and change to NumPy arrays for easier manipulation. The masks were padded with Zeros for the areas not masked to the original shape. In the last procedure, we need to handle the mask area overlapping problem. In the last procedure, we need to handle the mask area overlapping problem.

To solve this problem, the confidence scores of the objects in both bounding boxes would be compared. In the situation overlapping, one with higher confidence scores was used to mask the object. The overlapping area was eliminated from the one with smaller confidence scores.

4. EXPERIMENTATION AND RESULTS

We evaluate our algorithm datasets with different applications. It focuses on the application of vehicles detection and autonomous driving scenes. Each of both datasets consists of 5000 images.

In our experiment, YoloV4 model was used with pre-trained dataset. Decoder has been adapted to predict contour points. The models are trained end-to-end for schedule weighted factor λ. This helps the model to focus on regression loss in beginning of training to avoid self-intersecting polygons. Later, Deep sort algorithm is used along with the YOLOv4 to improve the segmentation accuracy and the pace of Object tracking and detection process involved before segmenting them into separate entities.

$$\lambda = \max(0.7822 + \frac{0.3429}{epoch}, 0.2)$$

Figure 4 – Scheduled Weight Factor

Table-1: Algorithm Comparsion Analysis

Experiment	AP_{50}	AP_{75}	Frame Rate
YOLOv4	62%	44%	20FPS
Tradition Algorithms	54%	38%	7FPS
YOLOv4-Deepsort	63.2%	44.8%	15FPS

Table 1 here illustrates the impact of our approach via the thesis we put forward with the Hybrid of alogorithms resulting in an increase of performance by 1.2% in Average Precision at 50% and 0.8% relative to YOLOv4 on Average Precision at 75%, being one of the latest algorithm involved in the tracking and detection process.

The model provides to have Frame rates to be lesser than Yolact encoded YOLOv4, however the accuracy is 2% more relative to it. For fair comparison, the encoder used in the architecture is DarkNet-53 which helps avoid misleading results.

Thus by the help of Cost function of DNN and further by the process of implementing the integration of Deep sort and YOLOv4 we end up in the nuch further enhanced Object tracking method as show in Fig 5, which is a snapshot of the tracking and detection after implementation of the algorithm.

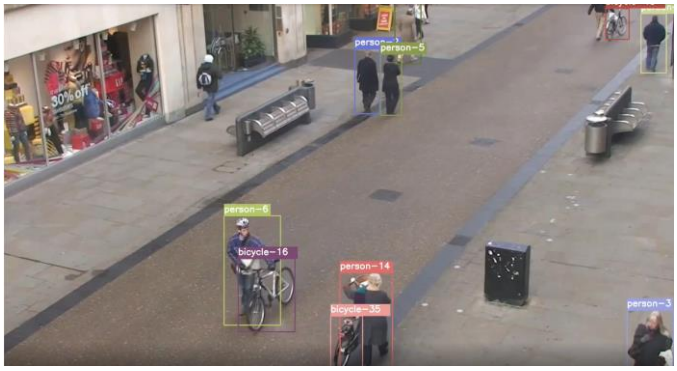

Figure 5 – Object detection and Tracking using YOLOv4 and Deepsort using Street Dataset

Figure 6 – Object detection and Tracking using YOLOv4 and Deepsort on Cars Dataset

5. REFERENCES

[1] Kato, N., Yamasaki, T., & Aizawa, K. (2019). Zero-shot semantic segmentation via variational mapping. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 0-0).

[2] Kalaivani, A., & Suguna, M. R. (2021). PERFORMANCE ANALYSIS OF VIDEO SHOT SEGMENTATION ALGORITHM. European Journal of Molecular & Clinical Medicine, 7(5), 1773-1779.

[3] Pan, S., Sun, S., Yang, L., & Duan, F. (2015, December). Video Shot segmentation algorithm based on SURF. In 2015 4th Int. Conference on Mechatronics, Materials, Chemistry and Computer Engineering. Atlantis Press

[4] Haroon, M., Baber, J., Ullah, I., Daudpota, S. M., Bakhtyar, M., & Devi, V. (2018). Video scene detection using compact bag of visual word models. Advances in Multimedia, 2018.

© 2021, www.IJARIIT.com ***All Rights Reserved***

[5] Haloi, P., Bordoloi, A. K., & Bhuyan, M. K. (2019, March). Unsupervised Broadcast News Video Shot Segmentation and Classification. In 2019 2nd International Conference on Innovations in Electronics, Signal Processing and Communication (IESC) (pp. 243-251). IEEE.

[6] Lu, X., Wang, W., Shen, J., Tai, Y. W., Crandall, D. J., & Hoi, S. C. (2020). Learning video object segmentation from unlabeled videos. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8960-8970).

[7] Liang-Chieh Chen et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: IEEE transactions on pattern analysis and machine intelligence 40.4 (2018), pp. 834–848.

[8] Zoran Zivkovic et al. "Improved adaptive Gaussian mixture model for background subtraction." In: ICPR (2). Citeseer. 2004, pp. 28–31.

[9] Zoran Zivkovic and Ferdinand Van Der Heijden. "Efficient adaptive density estimation per image pixel for the task of background subtraction". In: Pattern recognition letters 27.7 (2006), pp. 773–780.

[10] Kaiming He et al. "Mask r-cnn". In: Proceedings of the IEEE international conference on computer vision. 2017, pp. 2961–2969.

[11] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. "Learning representations by back-propagating errors". In: Cognitive modeling 5.3 (1988), p.1.

© 2021, www.IJARIIT.com ***All Rights Reserved***