Journal of Management & Entrepreneurship
ISSN 2229-5348

UGC Care Group I Journal
Vol-7 Issue-01 2018

HSV-based and Deep Learning-based Object Detection Algorithms for Recognizing Pedestrian Traffic Light Signals: A Comparative Study

[1]B.Naresh,Assistant Professor,
[2]E Krishna,Assistant Professor,
[3]Kalvala Sadanandam ,Associate Professor,
Department of CSE Engineering,
Nagole Institute of Technology and Science,
Kuntloor(V),Hayathnagar(M),Hyderabad,R.R.Dist.-501505.

Abstract—Small object detection has been a challenge in many image analysis applications. One such application is the ability to detect the status of Pedestrian Traffic Light (PTL) signal to allow decisions to be made by an intelligent system. The challenge is becoming more complex due to the increased complexity of urban environment, where objects of close similarity would confuse the detection mechanism. In this research, a study is carried out to compare two methods for the detection of small objects within large-sized images. The first method is a classical color-based segmentation approach while the second uses an intricate Deep Learning (DL) object detection algorithm. In the classical approach, objects within the selected range of Hue, Saturation and Value (HSV) composition are identified and extracted from the large-sized images. For DL approach, a Mask R-CNN was used where traffic light-like objects are identified by object instance segmentation process. From this research, it is shown that a two-tier approach, a hybrid HSV-DL model can detect the PTL signal directly and accurately from large-sized images in real-time on smart devices at an accuracy of 92.75%.

Keywords-deep learning, mask R-CNN, image segmentation, object detection, object classification

INTRODUCTION

Deep Learning, a subset of Machine Learning, within the field of artificial intelligence has been developed at an unprecedented rate. Complex tasks such as recognition and detection of objects are now achievable to a large extent due to the revolution of deep learning: neural networks that rely on automatic feature extraction through convolutional layers. The applications of this field extend from classifying different types of animals [1, 2] to autonomous vehicles [3, 4] as well as diagnosing stages of cancer [5].

However, despite the high success rate of the abovementioned applications as well as the progress of the deep learning field, the detection of small objects within a largesized image ranging from 4,096×3,072 pixels to 7,680×4,320 pixels in ultra-high-definition frame remains a challenge. One such example is the accurate detection and classification of pedestrian traffic light (PTL) signals. A typical 2D colored image consists of three channels – Red, Green and Blue (RGB), made up of pixels that range from 0 to 255 within each channel. These pixels make up the input layer of deep learning neural networks where with each progressing layer, it goes through feature extractions and non-linearity functions. The images are usually normalized and resized to a smaller dimension such as 300x300 pixels [SSD model] or 1,024x1,024 pixels [Faster R-CNN or R-FCN]. In the case of pedestrian traffic light signals which are relatively small objects, it may become unrecognizable if the source image is overly compressed, making the detection of PTL signal a near impossible task. In addition, the existence of objects similar to the shape or form and colors of PTL further complicate the detection process.

RELATED WORK

The commonly used approach is the classical color-based segmentation method, such as those shared in references [6, 7]. Reference [6] proposed a Traffic-Light Recognizer to support the visually impaired person using contour and color-based approach in order to identify potential Active Output Unit candidatures. Reference [7] proposed similar approach, the extraction algorithm is based on color segmentation and geometrical properties analysis while the recognition algorithm is based on SVM classification. These approaches yield fantastic results. However, different designs and forms of pedestrian traffic lights adopted in different countries may have impact on the robustness and accuracy of the detectors, making the deployment of such systems a challenge. Traffic lights that are partially occluded could also contribute to confusion in the detection process.

Currently there are machine learning and deep learning approaches deployed, such as the HSV-based analytic image processing and learning-based processing by [8] and Faster R-CNN-like models by [9]. Reference [9] addressed the challenge using a novel attention model based on a Faster RCNN algorithm. The locator and recognizer then uses another Faster R-CNN-like model. This has motivated us to explore the state-of-the-art deep learning approaches to provide a more robust solution that can be generalized for wider applications.

PROPOSED CONCEPT FOR PEDESTRIAN TRAFFIC LIGHT DETECTION

In this paper, we propose a novel method in addressing this problem, with the aim of addressing the issues raised in

the previous section and in making the detection more robust and efficient, as summarized in Fig. 1.

We suggest two different methods of segmentation with the objective of identifying regions of interest (ROI) of traffic light from a large-size image. The first method uses a classical Hue, Saturation and Value (HSV) based segmentation method, while the second uses Mask R-CNN instance segmentation. Both methods consist of two stages: the first stage analyses and identify ROI where there is potential presence of traffic light while the second stage uses a deep learning image classification network to classify the type of signals found in these ROIs, which may contain various traffic lights as well as background objects.

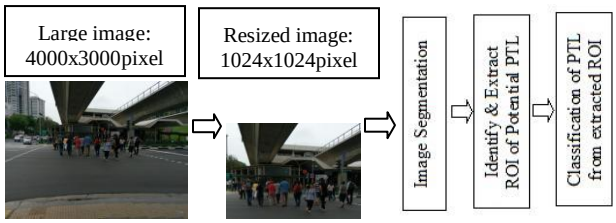

Figure 1. Proposed concept of PTL signal detection approach

HSV-BASED IMAGE SEGMENTATION

Segmentation Based on HSV

HSV color representation was designed in the 1970s by computer graphics researchers to more closely align with the way human vision perceives color-making attributes. Colors of each hue are arranged in a radical slice around a central axis of neutral colors that range from black to white. With these color concepts, desired colors can be filtered out from the image based on the range of HSV threshold values specified.

Implementation

Pedestrian traffic light consists of a Red Man to show that pedestrians are not allowed to cross a road while a Green Man is used to show that it is safe to cross a road. The colors within captured images are commonly expressed in the RGB color model [10]. In order to obtain the range of HSV values desired, the RGB values of the light signals are converted to HSV via the formulae given in Reference [11].

Using three different cameras with different camera specifications, 200 images were captured from various locations and in different lighting conditions for the purpose of HSV threshold selection. The cameras used are an 8 Megapixels (MP) Raspberry Pi Camera, HTC 10 mobile phone with a camera of 9.1 MP and a Samsung S9 mobile phone with a camera of 12MP. The images captured were taken in three different conditions: sunny skies, cloudy skies and when it is drizzling. Various perimeters were considered during image capturing to ensure a careful consideration for the HSV thresholds. The night scene was excluded in this study as it involves a different set of imaging techniques. Out of the 200 images captured, 100 were Green Man signals while the other 100 were Red Man. Based on the images captured, the range of HSV values were decided and tabulated in the table given below (see Table I).

TABLE I. HSVRANGE FOR RED AND GREEN MAN SEGMENTATION

Type	Range	Hue (°)	Saturation (%)	Value (%)
Red Man	Min.	329.76	70	55
	Max.	36.00	100	100
Green Man	Min.	158.00	73.6	19.6
	Max.	177.98	100	96.5

From the range of HSV values decided, the image is filtered: the entire image is converted into black aside from the objects that consists of colors within the HSV range specified. The filtered image then undergoes binary conversion. After which, the image is diluted and eroded for clarity, and labeled [12]. The features of the object, such as area, centroid and boundaries are extracted.

Results

Four hundred images were selected as test images for this research; 200 images containing Red Man signals while the other 200 are Green Man signals. In order to ensure a fair selection for the test images, they were captured from various locations in Singapore at different view distances of PTL. These distances range from one lane to seven car lanes, producing varying size of PTL in a typical street. An example of HSV segmentation is shown in Fig. 2.

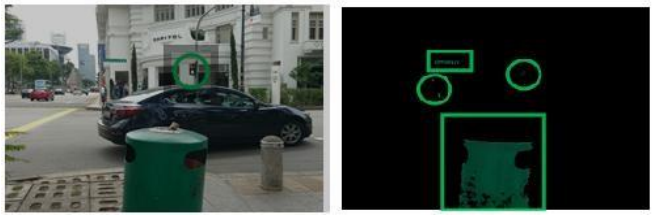

Figure 2. An image after going through HSV segmentation

Based on the HSV segmentation carried out on the test images, the size of the Red and Green Man ranges from 108 to 35 020 square pixels. The HSV segmentation process was successful in extracting most of the PTL signals from the test images, as tabulated in Table II.

RESULTS FOR EXTRACTION OFPEDESTRIANTRAFFIC SIGNALS FROM CLASSICAL HSVSEGMENTATION PROCESS

PTL Signal	No. of signal detected	Accuracy (%)
Green Man	196 / 200	98.00
Red Man	177 / 200	88.50
Total	*373 / 400*	*93.25*

Upon HSV segmentation, the centroid of each object is located and a boundary is created around the object. The image within the boundary is extracted into a new image. These images, examples shown in Fig. 3, are then sent for an image classifier (explained in section VI) trained for the classification of PTL signals.

Figure 3. New images created after HSV Segmentation, binary conversion, labelling and extraction for image classification.

MASK R-CNNOBJECT INSTANCE SEGMENTATION

Introduction to Mask R-CNN

Alternative solutions to the challenging task of object detection are the emerging deep learning algorithms. It ranges from fast and efficient algorithms such as Single Shot Detector (SSD) [13] and You Only Look Once (YOLO) [14], to the more advanced and accurate models of Region-Based CNN, such as Faster R-CNN [15].

The Mask R-CNN model was introduced by [16], evolving from the initial introduction of the Region-Based Convolutional Neural Network, or R-CNN. It is the most recent advancement of the family models and supports both object detection and instance segmentation. The architecture of the Mask R-CNN (Fig. 4) consists of two key baseline systems, namely the Faster R-CNN module and the Instance Segmentation module. Within the Faster R-CNN module, it consists of a ResNet-101 [17] feature extraction module, which is a 101-layer deep convolutional neural network, and an RPN (Regional Proposal Network). The ResNet-101 serves to extract the key features of object to be detected, pass them into the RPN to identify proposed regions where the object lies. The ROI Align module takes the object proposal and divides it into a certain number of bins where the data points are sampled and re-computed using bilinear interpolation. A Feature Pyramid Network (FPN) [18] is used to process the transformed features and to produce the final mask for the instance segmentation.

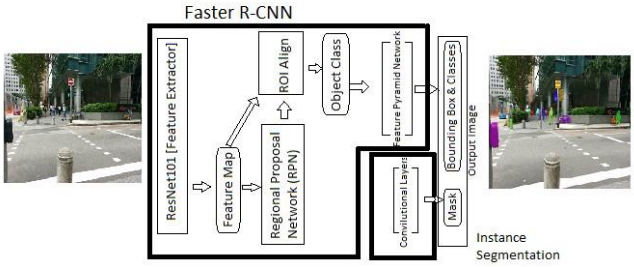

Figure 4. Mask R-CNN architecture

Implementation

There are several implementations of the Mask R-CNN, one of which is a fully tested open-source implementation [19], and has been used successfully in many implementation and applications such as life sciences, automation and city imaging. The codes were adapted for the PTL detection. The instance segmentation was simplified to only look for traffic related objects to improve the efficiency. These include the following object classes: pedestrian, car, motorcycle, bus, truck and traffic light. These objects are selected mainly to provide a context-aware solution in addition to a pure traffic light detection task.

C.Results

Based on the above implementation, the Mask R-CNN Segmentation was carried out on the same 400 test images used in the HSV-based Segmentation method mentioned previously. The results of the segmentation and extraction of PTL are tabulated in Table III.

RESULTS FOR EXTRACTION OF PEDESTRIANTRAFFIC SIGNALS FROM MASK R-CNNOBJECT INSTANCE SEGMENTATION

PTL Signal	PTL Signal detected	Accuracy (%)
Green Man	168 / 200	84.00
Red Man	191 / 200	95.50
Total	359 / 400	89.75

Similar to the previous implementation, the segmented objects are then sent to an image classifier.

CLASSIFICATION OF IMAGES

The segmented images contain very specific info but of vast variation in position and sizes. The CNN-based classification is deemed appropriate for this task.

Classification Using CNN

The evolution and highlight of deep learning are due to the creation of convolutional neural networks (CNNs). A CNN is a form of neural network that relies heavily on feature extraction layers called convolutional layers. Convolutional layers extract features from an image automatically based on filter size and number of filters that are determined by the neural net designer. The convolutional layer convolves the image input by sliding the filters along the input vertically and horizontally. As the convolution filter slides across the image input, the dot product of the weights and input (pixel values of image input) are computed, and a bias term is then added.

There are various parameters within the convolutional layer such as the type of padding used and the value of stride. Different weight initializers used such as Glorot [22], narrow-normal and zeros within the convolutional layer would also affect the accuracy of the classification network.

It is challenging to develop a robust classifier from scratch as it requires large dataset of training images and computing resources. Transfer learning, provides an alternative, where optimized weights from pre-trained networks are used as a starting point to cater to new networks for slightly different classification, such as the classification of PTL signals. With transfer learning, significantly fewer training iterations are needed and the weights used are already optimized to train new features of the segmented images. This in turn leads to shorter training time and computing resources required for image classifier training.

B.Implementation

AlexNet is the name of a convolutional neural network that was designed by Alex Krizhevsky [23]. It has a depth of eight layers. Despite the small number of layers, it has a

parameter of 61 million. The image input size is 227 by 227 with a channel size of 3 (colored images). Transfer learning using AlexNet was carried out using a **stochastic gradient descent with momentum** (SGDM) optimizer on **seven** classes: greenMan, redMan, digits, carRed, carGreen, nonRed and nonGreen. In deep learning neural networks, there are two common scenarios: a neural network that has over fitted to the training data due to the large depth of layers and a neural network that is not able to capture sufficient features for an accurate classification due to insufficient number of layers. Therefore, to avoid the issue of overfitting and still maintaining a network depth that is capable of capturing enough features of the Red and Green Man, regularization was used to add a coefficient into the error function calculation. This coefficient penalizes peaky weight vectors and ensure that the network is using all its input for training rather than using only a few inputs. Shuffling was also carried out at every epoch to ensure wide distribution during training to avoid overfitting.

Five thousand PTL images were captured to form the training images. Within each image captured, there are varying number of traffic lights ranging from one to five traffic lights. These traffic lights may be PTLs or car traffic lights. The number of images used for training are tabulated in Table IV.

Results

At the end of training, the CNN for PTL light signal image classification achieved a validation accuracy of **98.61%** with a validation loss of **0.0546**. The network was then tested against the **400 test images** and the results are tabulated below for both implementations (see Tables V, VI, VII and VIII).

TABLE IV. NUMBER OF TRAINING IMAGES USED FOR NETWORK TRAINING

Name of Cla ss:	No. of Imag es
Red Ma n	44 15
Green Ma n	47 37
Green Car Sign al	13 78
Red Car Sign al	12 04
Other Red Colored Obje cts	94 84
Other Gre en Colored Obje cts	25 30
Digits on the Traffic Lig ht	17 56

TABLE V. CONFUSION MATRIX FOR HSVSEGMENTATION METHOD ON THE TEST IMAGE SET

n = 3659		True Lab			
		Red	Green	Noise[a]	
Predicted Labels	Red	175	0	2	177
	Green	0	196	0	196
	Noise	2	0	3284	3286
		177	196	3286	3659

a. Other green / red objects used for training are classified as Noise

TABLE VI. PRECISION &RECALL FOR HSVSEGMENTATION METHOD

Color	Precision	Recall
Red	175 / 177 = 0.989	175 / 177 = 0.989
Green	196 / 196 = 1	196 / 196 = 1

TABLE VII. CONFUSION MATRIX FOR MASK R-CNNOBJECT INSTANCE SEGMENTATION ON THE TEST IMAGE SET

n = 1069		True s Label			
		Red	Green	Noise[a]	
Predicted Labels	Red	189	0	2	191
	Green	0	168	8	176
	Noise	2	0	700	702
		191	168	710	1069

a. Other green / red objects used for training are classified as Noise.

TABLE VIII. PRECISION &RECALL FOR MASK R-CNNOBJECT INSTANCE SEGMENTATION

Color	Precision	Recall
Red	189 / 191 = 0.989	189 / 191 = 0.989
Green	168 / 176 = 0.955	168 / 168 = 1.000

The neural network trained can classify both Red and Green Man accurately even when it is partially blocked by external and unforeseen objects (e.g. umbrellas, people and busses). The misclassifications can be resolved by placing a minimum confidence in actual real-life application: such as a 90% or higher in confidence level.

Overall, from the Precision-Recall calculations carried out, the image classification performed for the **HSV-based Segmentation method** produces better result and it achieves a **higher precision and recall** value.

RESULTS &DISCUSSION

Considering the accuracy of object segmentation and classification through the two-stage approach, the final (average) detection results is tabulated in Table IX. The issues on robustness, accuracy and ease of real-time implementation are discussed in the following sub-sections.

TABLE IX. OVERALL RESULTS FOR BOTH IMPLEMENTATIONS (AFTER SEGMENTATION AND CLASSIFICATION)

Methods	No. of PTL Successfully Segmented & Classified	Accuracy for G and R Signal	Average Accuracy (%)
HSV-based + CNN Classifier	**G:** 196 /200 **R:** 175 / 200	**G:** 98% **R:** 87.5%	**92.75%**
Mask R-CNN + CNN Classifier	**G:** 168 / 200 **R:** 189 / 200	**G:** 84% **R:** 94.5%	**89.25%**

Robustness of Implementation

Both implementations (HSV-based and Mask RCNN segmentation) ensures that even in conditions where the traffic light signals are blocked to a large degree, the traffic light signals can still be detected, some of which are shown in Fig. 5. This is because the implementation does not entirely depend on geometrical properties analysis such as form and contour of a PTL, but rather the features extracted by the CNN classifier, making the approach more robust than some of the existing form and contour based solutions.

Figure 5. Sample of partially blocked images classified correctly

Complexity and Accuracy

Based on the overall results obtained, the HSV-based approach was more accurate than the Mask RCNN model. The HSV composition of the green and red lights are rather unique from the traffic and road environment, with the exception that the red light might be confused as certain objects such as red plastic bags, red car signals and car headlamps. It has the advantages of ease of implementation, less complexity in addition to higher accuracy.

Mask Regional-CNN method, on the other hand, is known to be complex and requires more computing resources. It appears to be slightly less accurate as compared to the HSV approach. This is because the DL model was built for general object detection, it is able to detect more than 80 objects from a given scene. Due to this, it has the advantages of providing contextual information to the scene, such as vehicles detected on the road would give a different scenario from having pedestrian detected on the road.

Implementation for Real Time Detection

The HSV-based algorithm has been implemented successfully on a MS Surface tablet, capable of detecting PTL up to 60FPS real-time video stream. On the other hand, Mask R-CNN is not able to perform real-time detection using a typical smart device driven only by a central processing unit (CPU). It could be possible in the near future by integrating cloud-based solution and 5G mobile network technologies.

CONCLUSION

The HSV-based Segmentation with CNN Classifier approach is a preferred solution as it provides high accuracy in PTL signal detection, and is able to provide real-time detection, which is a critical criterion for practical applications. On the other hand, while Mask R-CNN Object Instance Segmentation algorithm may be complex, its complexity allows detection of multiple object classes that provides the user with more contextual information of the given situation. Furthermore, the Mask R-CNN approach is not limited by objects that may not have a distinct color such as the PTL signal. We believe that the constraint faced by complex detection algorithms can be overcome in the near future with the development of GPU enabled devices such as the NVIDIA Jetson Nano and the availability of 5G Network technologies.

ACKNOWLEDGEMENT

The authors wish to thank Tote Board-Enabling Lives Initiative and Nanyang Polytechnic for providing the resources to support the conduct of this research.

REFERENCES

[1] ***Chen, G., Han, T. X., He, Z., Kays, R., & Forrester, T. (2014). Deepconvolutional neural network based species recognition for wild animal monitoring (pp. 858–862). IEEE International Conference on Image Processing (ICIP).***

[2] ***Gomez Villa, A., Salazar, A., & Vargas, F. (2017). Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. Ecological Informatics, 41, 24–32.***

[3] ***Mohammed Al-Qizwini, Iman Barjasteh, HothaifaAlQassab, and Hayder Radha. Deep learning algorithm for autonomous driving using googlenet. In Intelligent Vehicles Symposium (IV), 2017 IEEE, pages 89–96. IEEE, 2017.***

[4] ***Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In Proceedings of the IEEE International Conference on Computer Vision, pages 2722–2730, 2015. 2***

[5] ***Olliverre N., Yang G., Slabaugh G., Reyes-Aldasoro C.C., Alonso E. International Workshop on Simulation and Synthesis in Medical Imaging. Springer; Cham, Switzerland: 2018. Generating Magnetic Resonance Spectroscopy Imaging Data of Brain Tumours from Linear, Non-linear and Deep Learning Models; pp. 130–138.***

[6] ***SergioMascetti, Dragan Ahmetovic, Andrea Gerino, Cristian Bernareggi, Mario Busso, Alessandro Rizzi. Robust traffic lights detection on mobile devices for pedestrians with visual impairment; Computer Vision and Image Understanding 2016, DOI:10.1016/j.cviu.2015.11.017***

[7] ***Ruiqi Cheng, Kaiwei Wang, Kailun Yang, Ningbo Long, Jian Bai, Dong Liu; Real-time pedestrian crossing lights detection algorithm for the visually impaired; Multimedia Tools and Applications; August 2018, Volume 77, Issue 16, pp 20651–20671.***

[8] ***R. de Charette , F. Nashashibi , Traffic light recognition using image processing compared to learning processes, in: Proceedings of the 22nd International Con- ference on Intelligent Robots and Systems, IEEE, 2009, pp. 333–338 .***

[9] ***Yifan Lu1, Jiaming Lu1, Songhai Zhang1, and Peter Hall; Traffic signal detection and classification in street views using an attention model: Computational Visual Media, Vol. 4, No. 3, 2018, 253–266.***

[10] ***"Analog and Digital Images," Principles of Remote Sensing - Centre for Remote Imaging, Sensing and Processing, CRISP, 2001. [Online]. Available: https://crisp.nus.edu.sg/~research/tutorial/image.htm. [Accessed: 24-Sep-2019].***

[11] ***"RGB to HSV conversion | color conversion", Rapidtables.com, 2019. [Online]. Available: https://www.rapidtables.com/convert/color/rgbto-hsv.html. [Accessed: 27- Sep- 2019].***

[12] ***Haralick, Robert M., and Linda G. Shapiro, Computer and Robot Vision, Volume I, Addison-Wesley, 1992, pp. 28-48***

[13] ***Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single Shot Multibox Detector. In European conference on computer vision, pages 21–37. Springer, 2016.***

[14] ***Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.***

[15] ***Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster RCNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015.***

[16] ***He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.***

[17] ***He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.***

[18] ***Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017; pp. 936–944.***

[19] ***Microsoft COCO Dataset, http://cocodataset.org/#home***

[22] ***Glorot, Xavier, and YoshuaBengio. "Understanding the difficulty of training deep feedforward neural networks." In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249-256. 2010.***

[23] ***Gershgorn, D. (2018). Rise of Alexnet: The inside story of how AI got good enough to dominate Silicon Valley. QUARTZ.***