Multi-Class Vehicle and Pedestrian Classification Using Convolutional Neural Network for Traffic Flow and Congestion Composition Analysis

Robert Kerwin C. Billones, Alexis M. Fillone, and Elmer P. Dadios

***Abstract*—This paper presents the development of a multi-class vehicle and pedestrian detection and classification using convolutional neural network (CNN) for the analysis of traffic flow and congestion. The study focused on analyzing the traffic flow and volume at different time intervals in a microscopic scale traffic network by decomposing it into eight separate classes of vehicles and pedestrians. Traffic videos in low altitude view T-type intersection (with pedestrian lane and yellow box area), medium altitude view bus stop area, and high altitude view wide intersection are used in the analysis of different traffic flow and congestion scenarios. The CNN model used have a 78.41% training accuracy with 0.6570 loss, and 73.83% validation accuracy with 0.7083 loss for the eight output multi-object classification. The results also showed how each component (class) contributes to the overall road traffic. Private cars constitute about 55-70% of the total traffic volume at any given time, while public utility vehicles (PUVs, jeepneys, buses) only takes approximately 15%. The showed that the implementation of CNN for classification is effective.**

***Keywords*—convolutional neural network, computer vision, multi-class object classification, traffic flow and congestion analysis**

I. Introduction

Traffic flow can be analyzed in a macroscopic and microscopic scale. In macroscopic scale, traffic flow can be viewed as a compressible and continuous fluid movement with static behavior as it flows through the medium [1] [2] [3]. In microscopic scale, road traffic elementary particles can be observed in analyzing traffic flow and congestion. It takes into consideration individual behavior and interactions of traffic participants, vehicles and pedestrians, in traffic networks [4] [5]. These behaviors and interactions can be analyzed using computer vision [6] [7] [8] [9], and machine learning techniques [10]. The rich information from traffic videos can be used to observe vehicle flow [11], speed, and density. It can also provide information on pedestrian movements, and how it affects the over-all traffic situation. This study focused on vehicle and pedestrian composition in different traffic scenarios to analyze traffic flow and congestion. Traffic video data sets for low altitude view T-type intersection (with pedestrian lane, and yellow box area), medium altitude view bus stop area, and high altitude view wide intersection are used for analysis. Using a Haar cascade classifier, vehicles and pedestrians can be detected in traffic videos. The convolutional neural network (CNN) is used to further classify these vehicles and pedestrians into different sub-classes. Usingia this method, vehicle and pedestrian composition in traffic flow and congestion can be observed and analyzed.

De La Salle University, 2401 Taft Avenue, Manila 0922, Philippines (email: robert.billones@dlsu.edu.ph, alexis.fillone@dlsu.edu.ph, elmer.dadios@dlsu.edu.ph)

II. Haar Cascade Classifier

A Haar feature sets up a detection window where it calculates the sum of pixel intensities in each group of neighboring pixel regions. A category is assigned to the difference between those sums to label the sub-sections of an image [12]. The cascade classifier uses a detection window and slides it over the image to detect features. Each stage in the cascade contains the list of weak learners. The integral image concept of Viola and Jones [13] are used in Haar features computation [14] [15] [16], see eq.1.

Copyright © 2020 by De La Salle University

$$ii(x,y) = \sum_{x \le x', y \le y'} i(x',y') \quad (1)$$

Where $i(x,y)$ is the original image, and $ii(x,y)$ is the integral image. Haar feature windows computation at different image scale can be achieved using eq. 2 and 3.

$$s(x,y) = s(x,y-1) + i(x,y) \quad (2)$$
$$ii(x,y) = ii(x-1,y) + s(x,y) \quad (3)$$

Where $i(x,y)$ represents the cumulative sum of the original image.

III. Convolutional Neural Network

CNN is a multi-layer neural network architecture that takes advantage of spatial relationships found in images and videos. Information propagates through network of layers are digitally filtered, at each layer, to obtain relevant features. This technique allows a neuron to access basic features, such as corners and edges, and have invariance to scale, shift, and rotation [17] [18] [19] [20]. The three key layers of a CNN are the convolutional layer (CONV), the pooling layer (POOL), and the fully-connected layer (FC). The CONV layer is connected spatially (width and height) to a small portion of the input layer, but to full depth, and computes for each dot products, see Fig. 1.

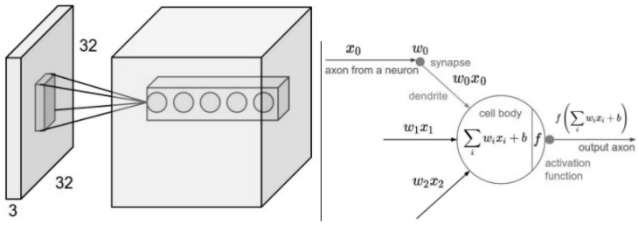

Fig. 1. An example of CNN convolutional layer [21]

The POOL layer is usually inserted between consecutive CONV layers and perform spatial down-sampling to reduce the number of representations. It is also used to avoid overfitting. The POOL layer receives an input of size W1×H1×D1 and produces an output of W2×H2×D2 by using spatial extent (F), and stride (S) parameters, see eq. 4-6.

$$W2=(W1-F)/S+1 \quad (4)$$
$$H2=(H1-F)/S+1 \quad (5)$$
$$D2=D1 \quad (6)$$

Each neuron in the FC layer is connected to all the neurons in the previous layer. The computation of class scores is done in this layer [21].

IV. Methodology

A. Traffic Video Data Sets

Traffic video data sets for low altitude view T-type intersection (with pedestrian lane, and yellow box area), medium altitude view bus stop area, and high altitude view wide intersection are used in this study, see Table 1. The altitude (H) of camera views are categorized into low (3m to 5m), medium (5m to 10m), and high (10m and above) altitudes. Fig. 2 shows the typical road camera setup. Table 2 lists the details regarding the traffic video data sets such as duration, resolution, and frames per second (FPS).

TABLE 1
Traffic Video Data Sets: Camera Positioning (Altitude), Purpose (Calibration/Testing), and Description

Dataset	Altitude	Purpose	Description
DS0-LS1	Low	Calibration	Pedestrian lane
DS0-LS2	Low	Calibration	T-intersection
DS0-LS3	Low	Calibration	T-intersection
DS0-TA1	Low	Calibration	Normal road
DS0-TA2	Low	Calibration	Yellow box area
DS0-TA3	Low	Calibration	Normal road
DS3-1	High	Calibration	Wide intersection
DS4-1	Medium	Calibration	Bus stop area (day-time)
DS4-3	Medium	Calibration	Bus stop area (night-time)

TABLE 2
Traffic Video Data Sets: Duration, Resolution, and FPS

Dataset	Video	Duration	Resolution	FPS
DS0-LS1	CA1_R_LS1	00:12:06	2560x1440	25
DS0-LS2	CA1_R_LS2	00:12:05	2048x1536	25
DS0-LS3	CA1_R_LS3	00:12:02	2304x1296	25
DS0-TA1	CA1_R_TA1	00:12:02	1920x1080	25
DS0-TA2	CA1_R_TA2	00:12:06	1280x720	25
DS0-TA3	CA1_R_TA3	00:12:06	2560x1440	25
DS3-1	MMDA_A_NO1	03:59:49	800x452	25
DS4-1	MMDA_A_SH1	06:00:00	1280x720	12
DS4-3	MMDA_A_SH3	06:00:00	1280x720	12

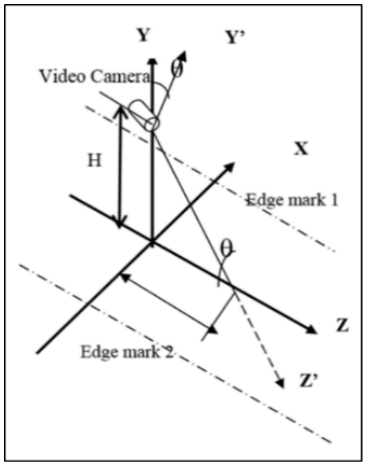

Fig. 2. Camera setup [22]

B. Haar Cascade Parameters

In this study, values for Haar cascade parameters for vehicle and people detection were chosen for optimal detection, as shown in Table 3 and 4 respectively. The scale factor determines the reduction of image size at specific image scale, i.e., value of 1.1 reduces image size by 10%. Minimum neighbors determine the number of neighbors for each candidate to retain the detection. The minimum and maximum size detection window determine the possible object size that can be detected. Objects smaller or larger than these parameters are ignored, respectively.

TABLE 3
Haar Cascade Classifier for Vehicle Detection
(DS0, DS3-1, DS4-1, and DS4-3)

Dataset	scaleFactor	minNeighbors	minSize	maxSize
DS0-LS1	1.1	1	150	250
DS0-LS2	1.1	1	150	250
DS0-LS3	1.1	1	100	250
DS0-TA1	1.1	1	100	250
DS0-TA2	1.1	1	50	200
DS0-TA3	1.1	1	150	250
DS3-1	1.1	1	30	60
DS4-1	1.1	1	50	100
DS4-3	1.1	1	50	200

TABLE 4
Haar Cascade Classifier for People Detection
(DS0, DS3-1, DS4-1, and DS4-3)

Dataset	scaleFactor	minNeigbors	minSize	maxSize
DS0-LS1	1.1	2	50	150
DS0-LS2	1.1	2	50	150
DS0-LS3	1.1	2	50	200
DS0-TA1	1.1	1	50	100
DS0-TA2	1.1	1	50	150
DS0-TA3	1.1	1	50	150
DS3-1	1.1	1	10	20
DS4-1	1.1	1	10	30
DS4-3	1.1	1	50	150

Haar cascade classifier (vehicle and people) detection accuracy used in this study can be computed using the eq. 7 [23]:

$$Accuracy = \frac{(TP+TN)}{N} x100\% \quad (7)$$

Where TP = true positive, TN = true negative, and N = total number of detected objects.

C. CNN Model, Training, and Validation

The CNN architecture model used in this study is a Keras sequential model [24]. It uses four layers of 2D convolutional layer which creates a convolutional kernel that is convolved with the input layer to produce a tensor of outputs. The first three CONV layers have 32 filters, and 3x3 kernel size. The last CONV layer has 64 filters, and 3x3 kernel size. Each of these layers used a rectified linear unit (relu) activation, and 2D max pooling layer with 2x2 pool size. The pool size down scale it by a factor of 2 in the vertical and horizontal spatial dimension. After the 2D CONV layers, the core layer used a dropout of 0.5, sigmoid activation function, and a densely connected neural network layer of eight (8) output classes. The CNN architecture used a categorical cross entropy loss function, RMSprop optimizer, and accuracy for metrics.

Image data set derived from Haar cascade classifier are used for training, validation, and testing. The CNN have eight (8) output classes which are: private cars, motorcycles, PUVs, jeepneys, buses, trucks, people, and indistinguishable images. Each class have 1000 training, and 200 validation, except for class 4 (jeepneys). This class (jeepneys) have 370 training, and 100 validation images. A total of 7370 training, and 1500 validation images are used to train and validate the CNN model. After training and validation, h5 files for the model and weights were generated.

V. EXPERIMENT AND RESULTS

A. Generating Vehicle Classification Image Data Set using Haar Cascades

A pre-trained Haar cascade classifier for vehicle detection was used to generate an image training data set for localized vehicle classification. After Haar cascade detection, the detected objects were sorted manually into different classes, as show in Table 5 and 6. Fig. 3 and 4 show the sample generated vehicle images for DS0, DS3-1, DS4-1, and DS4-3. The valid classes of vehicles were private cars, motorcycles, public utility vehicles (PUVs), jeepneys, buses, and trucks. Bicycles and tricycles were included in the motorcycle class. Taxis and public utility vans were included in the PUV class. Coaster bus and school bus are considered in the bus class. Lastly, trailer trucks, service trucks, delivery trucks, and armored trucks were considered in the truck class. Detected objects other than vehicles were considered invalid classes of vehicles. This may include people and indistinguishable objects. As shown in Fig. 5, images with high occlusion, non-vehicle, non-people, and multiple objects were categorized into the indistinguishable class. Valid vehicle classes are considered true positives, while invalid classes are false positives. True negatives were set to zero in the performance computation. The performance of vehicle detection using Haar cascade classifier for DS0, DS3-1, DS4-1, and DS4-3 are shown in Table 7. Vehicle detection accuracy are low for DS0 and DS4-1, while DS3-1 and DS4-3 have an acceptable detection accuracy.

TABLE 5
GENERATED VEHICLE CLASSIFICATION IMAGE DATA SET (DS0)

Object Class	DS0-LS1	DS0-LS2	DS0-LS3	DS0-TA1	DS0-TA2	DS0-TA3
Private cars	38	626	120	51	20	34
Motorcycles	1	129	3	16	5	9
Public utility vehicles	20	85	18	21	9	54
Jeepneys	2	1	8	6	0	25
Buses	1	8	2	0	0	0
Trucks	0	2	0	0	0	0
People	1	0	0	0	0	0
Indistinguishable	59	421	128	22	13	16
Total No. of Detected Objects	**122**	**1272**	**279**	116	47	138

TABLE 6
GENERATED VEHICLE CLASSIFICATION IMAGE DATA SET (DS3-1, DS4-1, AND DS4-3)

Object Class	DS3-1	DS4-1	DS4-3
Private cars	12160	1156	111
Motorcycles	1676	3	0
Public utility vehicles	1903	81	9
Jeepneys	432	0	0
Buses	92	4217	1200
Trucks	1801	36	4
People	89	22	0
Indistinguishable	4473	4686	301
Total No. of Detected Objects	22626	10201	1625

TABLE 7
PERFORMANCE OF VEHICLE DETECTION USING HAAR CASCADE CLASSIFIER

Performance Metrics	DS0	DS3-1	DS4-1	DS4-3
True Positive (TP)	1314	18064	5493	1324
False Positive (FP)	660	4562	4708	301
Total No. of Detected Objects (*N*)	1974	22626	10201	1625
Accuracy (%)	66.57	79.84	53.85	81.48

Fig. 3. Sample generated vehicle classification images (DS0)

Fig. 4. Sample generated vehicle classification images (DS3-1, DS4-1, and DS4-3)

Fig. 5. Sample indistinguishable vehicle images (DS0, DS3-1, DS4-1, and DS4-3)

B. Generating People Image Data Set using Haar Cascades

A pre-trained Haar cascade classifier for people detection was used to generate an image training data set for localized people classification. After Haar cascade detection, the detected objects were sorted manually into different classes, as show in Tables 8 and 9. The sample generated people images for DS0, DS3-1, DS4-1, and DS4-3 is shown in Fig. 6. The valid classes of people are people riding motorcycles/ bicycles/tricycles and pedestrians. Detected objects other than people were considered invalid classes of people, such as vehicles and indistinguishable objects. True negatives were also set to zero in the performance computation. The performance of people detection using Haar cascade classifier for DS0, DS3-1, DS4-1, and DS4-3 are presented in Table 10. People detection accuracy were low for all data sets because of high number of indistinguishable objects detected by the classifier.

TABLE 8
GENERATED PEOPLE CLASSIFICATION IMAGE DATA SET (DS0)

Object Class	DS0-LS1	DS0-LS2	DS0-LS3	DS0-TA1	DS0-TA2	DS0-TA3
Motorcycles	1	206	1	5	5	34
People	16	151	6	10	66	248
Indistinguishable	67	146	39	48	24	119
Total No. of Detected Objects	**84**	**503**	**46**	63	95	401

TABLE 9
GENERATED PEOPLE CLASSIFICATION IMAGE DATA SET (DS3-1, DS4-1, AND DS4-3)

Object Class	DS3-1	DS4-1	DS4-3
Private cars	0	6	43
Motorcycles	2061	865	25
Buses	0	29	77
People	1000	18472	1478
Indistinguishable	7349	15631	844
Total No. of Detected Objects	**10410**	**35003**	**2467**

TABLE 10
PERFORMANCE OF PEOPLE DETECTION USING HAAR CASCADE CLASSIFIER

Performance Metrics	DS0	DS3-1	DS4-1	DS4-3
True Positive (TP)	749	3061	19337	1503
False Positive (FP)	443	7349	15666	964
Total No. of Detected Objects (N)	1192	10410	35003	2467
Accuracy (%)	62.84	29.40	55.24	60.92

Fig. 6. Sample generated people images (DS0, DS3-1, DS4-1, and DS4-3)

C. Multi-Class Object Classification using Convolutional Neural Network

The CNN model earlier have 78.41% training accuracy with 0.6570 loss, and 73.83% validation accuracy with 0.7083 loss for the eight output multi-class object classification. The training and validation simulation time is 56 minutes and 36 seconds with 13 epochs, see Table 11.

TABLE 11
CNN TRAINING AND VALIDATION RESULTS

Epoch	Training		Validation	
	Accuracy	**Loss**	**Accuracy**	**Loss**
0	0.00%	0.0000	0.00%	0.0000
1	35.35%	1.5897	62.35%	1.1172
2	55.68%	1.1700	67.91%	0.9225
3	65.56%	0.9688	68.22%	0.8080
4	68.29%	0.8676	68.22%	0.8288
5	71.97%	0.7996	73.52%	0.7467
6	72.42%	0.7643	73.83%	0.6983
7	75.06%	0.7147	71.03%	0.7258
8	76.24%	0.6934	73.60%	0.7915
9	76.24%	0.6993	72.27%	0.8377
10	78.24%	0.6707	75.00%	0.7503
11	77.95%	0.6642	71.88%	0.7835
12	78.33%	0.6589	70.56%	0.7719
13	**78.41%**	**0.6570**	**73.83%**	**0.7083**

D. Multi-Class Vehicle Traffic Flow Composition Analysis

The vehicle traffic composition analysis used the generated DS3-1 data, see Table 12. This data set was chosen because it had the highest total number of detected objects (N=22626) with 79.84% accuracy and have enough data points (4-hr video length). Only true positive objects (TP=18064) was considered for analysis. Traffic congestion build-up is usually attributed to high volume of vehicles in a road network that exceeds its road capacity. Other road traffic parameters, such as vehicle speed and traffic signalization can also affect traffic congestion. In this study, the effects of vehicle composition in traffic congestion build-up was analyzed. Table 13 shows vehicle composition in percentage while Fig. 7 shows the visual representation of the traffic composition in 10-min time frame. This data shows how each component (class) contributes to the over-all road traffic. Private cars constitute about 55%-70% of the total traffic volume at any given time, while public utility vehicles (PUVs, jeepneys, buses) only takes approximately 15%. Motorcycles usually have 6% up to 20% (peak), and trucks have 2% up to 25% (peak).

TABLE 12
MULTI-CLASS VEHICLE TRAFFIC COMPOSITION DATA IN 10-MIN TIME FRAME

Time frame (10-min)	Total TP objects	Private Car	Motor cycle	PUV	Jeep	Bus	Truck
1	748	487	47	70	29	4	111
2	571	360	73	77	9	0	52
3	691	479	72	76	8	3	53
4	746	515	58	80	35	2	56
5	607	333	42	57	21	1	153
6	576	343	45	121	14	4	49
7	758	558	56	96	13	6	29
8	572	388	51	59	8	7	59
9	658	398	106	103	9	0	42
10	778	519	154	72	12	2	19
11	666	434	100	67	13	21	31
12	529	371	74	43	4	3	34
13	572	351	94	68	6	1	52
14	700	463	68	59	42	2	66
15	557	376	73	52	8	2	46
16	671	462	70	56	9	0	74
17	862	575	63	93	57	10	64
18	835	602	65	92	35	2	39
19	935	659	57	84	5	2	128
20	881	632	54	89	8	0	98
21	928	611	99	85	13	1	119
22	997	666	63	100	43	0	125
23	1079	738	50	102	7	18	164
24	1147	840	42	102	24	1	138
TOTAL	**18064**	**12160**	**1676**	1903	432	92	1801

TABLE 13
Percentage of Vehicle Class Composition in 10-Min Time Frame

Time frame (10-min)	Private Car	Motor cycle	PUV	Jeep	Bus	Truck
1	65.11%	6.28%	9.36%	3.88%	0.53%	14.84%
2	63.05%	12.78%	13.49%	1.58%	0.00%	9.11%
3	69.32%	10.42%	11.00%	1.16%	0.43%	7.67%
4	69.03%	7.77%	10.72%	4.69%	0.27%	7.51%
5	54.86%	6.92%	9.39%	3.46%	0.16%	25.21%
6	59.55%	7.81%	21.01%	2.43%	0.69%	8.51%
7	73.61%	7.39%	12.66%	1.72%	0.79%	3.83%
8	67.83%	8.92%	10.31%	1.40%	1.22%	10.31%
9	60.49%	16.11%	15.65%	1.37%	0.00%	6.38%
10	66.71%	19.79%	9.25%	1.54%	0.26%	2.44%
11	65.17%	15.02%	10.06%	1.95%	3.15%	4.65%
12	70.13%	13.99%	8.13%	0.76%	0.57%	6.43%
13	61.36%	16.43%	11.89%	1.05%	0.17%	9.09%
14	66.14%	9.71%	8.43%	6.00%	0.29%	9.43%
15	67.50%	13.11%	9.34%	1.44%	0.36%	8.26%
16	68.85%	10.43%	8.35%	1.34%	0.00%	11.03%
17	66.71%	7.31%	10.79%	6.61%	1.16%	7.42%
18	72.10%	7.78%	11.02%	4.19%	0.24%	4.67%
19	70.48%	6.10%	8.98%	0.53%	0.21%	13.69%
20	71.74%	6.13%	10.10%	0.91%	0.00%	11.12%
21	65.84%	10.67%	9.16%	1.40%	0.11%	12.82%
22	66.80%	6.32%	10.03%	4.31%	0.00%	12.54%
23	68.40%	4.63%	9.45%	0.65%	1.67%	15.20%
24	73.23%	3.66%	8.89%	2.09%	0.09%	12.03%

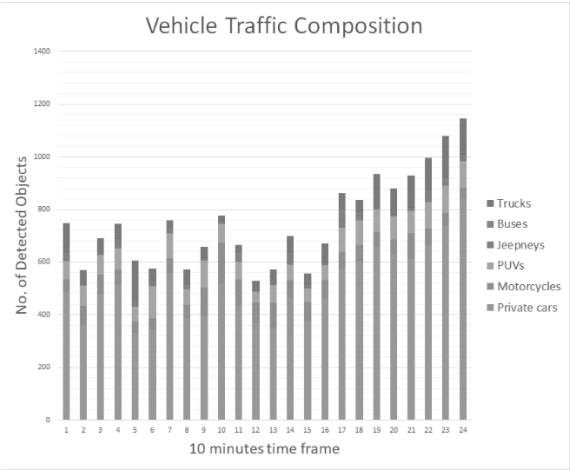

Fig. 7. Visual representation of multi-class vehicle traffic composition using DS3-1 data set

VI. Conclusion

The study aims to demonstrate a traffic monitoring and analysis method by treating vehicular and pedestrian movements as elementary particles that can be observed and analyzed individually, rather than viewing traffic flow and volume with the same behavior for all traffic participants. Using a traffic video data sets with different traffic scenarios, vehicle and people were first detected using Haar cascade classifier. Vehicle detection accuracy for DS0 is 66.57%, DS3-1 is 79.84%, DS4-1 is 53.85, and DS4-3 is 81.48%. The low accuracy for DS0 (T-type intersection, low altitude view), and DS4-1 (bus stop area, day time, medium altitude view) suggest the high concentration of activity with high number of occlusion and false positive detections for low-to-medium altitude camera views. DS3-1 (wide intersection, high altitude view) have good accuracy even with high concentration of activity because of small number of occlusion. DS4-3 (bus stop area, night time, medium altitude view) have good accuracy because of low concentration of activity during this time of day, and less number of occlusion and false positive detections. People detection accuracy for DS0 is 62.84%, DS3-1 is 29.40%, DS4-1 is 55.24%, and DS4-3 is 60.92%. These results suggest that for low-to-medium altitude camera view (DS0, DS4-1, and DS4-3), there are still high numbers of false positive detections. The poor performance for DS3-1 suggest that in high altitude camera views the people detection algorithm cannot discriminate enough between small vehicular movements and people. After detection, CNN is used to classify these detected objects into one of eight output classes (private cars, motorcycles, PUVS, jeepneys, buses, trucks, people, and indistinguishable images). CNN classification accuracy is 78.41% during training, and 73.83% during validation. Traffic flow and congestion can be separated into elementary particles (or individual classes) and analyzed these classes individually. Vehicle traffic composition for DS3-1 shows that at every 10-min window time frame, private cars constitute about 55% to 70% of the total traffic volume, while public utility vehicles (PUVs, jeepneys, buses) only takes approximately 15%. A continuous surge in volume of private cars caused a developing traffic congestion, as observed in the traffic volume the occlusion problem for low-to-medium altitude camera views should be addressed. High number of occlusions, as well as false positive detections, should be reduced in the detection stage. The classification algorithm should likewise be improved. Using the vehicle composition analysis presented in this study, a traffic congestion prediction or forecast algorithm can be developed as well.

ACKNOWLEDGMENTS

The authors highly appreciate the Department of Science and Technology - Philippine Council for Industry, Energy, and Emerging Technologies for Research and Development (DOST-PCIEERD) for providing funds for this research study.

REFERENCES:

[1] R. Boel and L. Mihaylova, "Modelling Freeway Networks by Hybrid Stochastic Models," *2004 IEEE Intelligent Vehicles Symposium ,* pp. 182-187, 2004.

[2] R. Danescu, F. Oniga and S. Nedevschi, "Modeling and Tracking the Driving Environment With a Particle-Based Occupancy Grid," *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS,* vol. 12, no. 4, pp. 1331-1342, 2011.

[3] W. Xiao-xiong, D. Gao-li, Y. Li-ping and W. Dan, "Decouple Analysis on Distributed Architecture of Urban Traffic System," *ICARCV 2006 ,* pp. 1-6, 2006.

[4] C. Zhang, H. Li, X. Wang and X. Yang, "Cross-scene Crowd Counting via Deep Convolutional Neural Networks," *IEEE Conference Publication,* pp. 833-841, 2015.

[5] J. Shao, K. Kang, C. C. Loy and X. Wang, "Deeply Learned Attributes for Crowded Scene Understanding," *IEEE Conference Publication,* pp. 4657-4666, 2015.

[6] N. Buch, S. A. Velastin and J. Orwell, "A Review of Computer Vision Techniques for the Analysis of Urban Traffic," *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS,* vol. 12, no. 3, p. 920, 2011.

[7] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng and H. Sawhne, "Video Event Recognition Using Concept Attributes," *IEEE Conference Publication,* pp. 339-346, 2013.

[8] Y. Zhu, N. M. Nayak and A. K. Roy-Chowdhury, "Context-Aware Activity Recognition and Anomaly Detection in Video," *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING,* vol. 7, no. 1, pp. 91-101, 2013.

[9] R. K. C. Billones, A. A. Bandala, E. Sybingco, L. A. G. Lim and E. P. Dadios, "Intelligent system architecture for a vision-based contactless apprehension of traffic violations," *2016 IEEE Region 10 Conference (TENCON),* pp. 1871 - 1874, 2016.

[10] R. K. C. Billones, A. A. Bandala, E. Sybingco, L. A. G. Lim, A. D. Fillone and E. P. Dadios, "Vehicle Detection and Tracking using Corner Feature Points and Artificial Neural Networks for a Vision-based Contactless Apprehension System," *Computing Conference 2017,* pp. 688-691, 2017.

[11] R. K. C. Billones, A. A. Bandala, L. A. G. Lim, E. Sybingco, A. D. Fillone and E. P. Dadios, "Microscopic Road Traffic Scene Analysis Using Computer Vision and Traffic Flow Modelling," *Journal of Advanced Computational Intelligence and Intelligent Informatics,* vol. 22, no. 5, pp. 1-6, 2018.

[12] M. Oualla, A. Sadiq and M. S., "A survey of Haar-Like feature representation," *International Conference in Multimedia Computing and Systems (ICMCS),* pp. 1101-1106, 2014.

[13] V. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition*, 2001.

[14] R. Lienhart and J. Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection," *IEEE ICIP 2002,* vol. 1, pp. 900-903, 2002.

[15] Nazeer, A. S, N. Omar, K. Jumari and M. Khalid, "Face detecting using artificial neural network approach," *First Asia International Conference Modelling & Simulation,* pp. 394-399, 2007.

[16] A. Mohamed, A. Issam, B. Mohamed and B. Abdellatif, "Real-time detection of vehicles using the haar-like features and artificial neuron networks," *The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015),* pp. 24-31, 2015.

[17] I. Arel, D. C. Rose and T. P. Karnowski, "Deep Machine Learning—A New Frontier in Artificial Intelligence Research," *IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE ,* pp. 13-18, 2010.

[18] F.-J. Huang and Y. LeCun, "Large-scale learning with SVM and convolutional nets for generic object categorization," *Proc. Computer Vision and Pattern Recognition Conf. (CVPR'06),* 2006.

[19] R. Girshick, "FastR-CNN," *2015 IEEE International Conference on Computer Vision,* vol. 8, no. 1, pp. 1440-1448, 2015.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper with Convolutions," *IEEE Conference Publication,* pp. 1-9, 2015.

[21] karpathy@cs.stanford.edu, "CS231n Convolutional Neural Networks for Visual Recognition," [Online]. Available: http://cs231n.github.io/convolutional-networks/. [Accessed 23 August 2017].

[22] J. Wu, Z. Liu, J. Li, C. Gu, M. Si and F. Tan, "An Algorithm for Automatic Vehicle Speed Detection using Video Camera," *Proceedings of 2009 4th International Conference on Computer Science & Education,* pp. 193-196, 2009.

[23] K. Qian, "Simple guide to confusion matrix terminology," Data School, 25 March 2014. [Online]. Available: http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/. [Accessed 26 March 2017].

[24] Keras, "Keras documentation," [Online]. Available: https://keras.io. [Accessed July 2018].