A Unified Framework on Reading Variable Size Traffic Sign Boards using Deep Neural Network

S. Roubil, Muhammad A. Hassan, Muhammad Usman Ghanni Khan
Al-Khwarizmi Institute of Computer Science UET, Lahore [1,2]
Computer Science Department UET, Lahore [3]
salman.roubil@kics.edu.pk, ahmed.hasan@kics.edu.pk, usman.ghani@kics.edu.pk

***Abstract*—With the explosive increase of vehicles in modern era and continuous interest in autonomous vehicles across the globe, need for automatic signs reading software has attained global attention. Recent research has focused on reading small sized sign boards whereas previous work was more related to understanding large sized sign boards. This work serves two-fold- identifying small or large sized sign boards within a single platform. This model makes use of recent advancements in deep learning paradigm entitle Inception V3, where this model is modified to cater needs of reading variable sizes of sign boards installed at traffic roads. This modified inception model is fed into Faster RCNN for better features extraction. RPN is applied for region of interest extraction. Later on, sign board is identified into three major classes, i.e. prohibited, warning or mandatory using fully connected layers of Neural Network. The proposed model achieves an accuracy of around 92.99% on a standard dataset.**

***Index Terms*—Deep Learning, Faster R-CNN, Traffic Sign Detection, Modified Inception v3.**

I. Introduction

Road accident are considered as one of the major problems; faced by human all over the world. According to World Health Organization (WHO)' statistics 1.35 Million people die each year in road crashes [1]. These numbers are continuously increasing with the infrastructure development of big cities. for many years, concerned authorities are implementing solutions to reduce these figures. Implemented solutions are typically applied as auto policies which focus on driver and traveler safety. It includes advance braking systems, precautionary alerts, maintaining distance with other vehicles and control over speed limits [2]. These Techniques are named as Advanced driver-assistance systems (ADAS) [3].

ADAS systems are designed to assist drivers while traveling. High death tolls in recent years made it compulsory for every auto manufacturer to invest their most of R&D budget in to ADAS systems [4]. These systems consist of many sub-modules [2]. Few of them directly control vehicle i.e. ABS systems, Cruise Control System, Radar Braking, while others assist drivers during drive. Most common and popular system for driver assistance is Navigation Guide [5]. Although accurate navigation guide systems are very hot topic in term of vehicle sales as well as for research but accidents are still happening.

This issue even gets worsen when it comes to sustainable cities of South Asian Countries. Specifically discussing Pakistan and India, it is observed that terrible road conditions with no proper sign boards increases difficulty level for drivers. That's why 93% of world's road fatalities happen in low- and middle-income countries. Other than death toll, 20 to 50 million people face non-fatal injuries which sometime cause lifelong disabilities [1]. World bank's research, conducted on data of 135 countries in first quarter of 2018, described that 10% reduction in deaths, caused by road accidents, will raise 3.6% over horizon of next 24 years [6].

To minimize the count of road accident a lot of necessary measures are urgently needed. Breaking down the techniques of driver and passenger' safety, it can be observed that road side sign boards for driver guidance are substantial [7]. The precautionary guidance provided by traffic sign boards, play vigorous role in assisting drivers, saving assets and safeguarding lives. But it is observed that driver frequently goes through difficulties in reading and understanding of sign boards when going on high speed, it occurs because driving itself involves close attention. These sign boards work as companion to driver who is continuously advising him to follow proper rules and driving habits. Due to its importance, researchers had been working on this topic since last 2 decades.

In early years of development regarding this issue, researchers had been focusing on shallow learning techniques to detect traffic sign boards [8]. But due to its limitation of feature extraction, it was not useful in different lighting conditions [9] as well as varying background scenes [10]. Another issue in those techniques was amount of data. Big data was hard to handle though extracting features using conventional computer vision. To cover that gap, deep learning [11] methods were adopted. It helped in gaining better accuracies for traffic sign board detection [12]. It is observed that research in area of traffic signs detection by deep learning doesn't perform well on small sign boards in whole scene. Its is because most of detection algorithms ignore small object features going through image convolution [13].

To minimize this problem, we proposed solution named "Begleiter". It assists driver by detecting roadside sign board then understands signs and alerts driver according to severity of sign. This work will discuss architecture of Begleiter using deep convolution neural networks. It is constructed by using Region Proposed Network (RPN) [14] with inception module. Our proposed research will help reducing number of accidents, which will eventually boost [15] GDP specially for developing countries. Our proposed research will help reducing gaps

978-1-7281-3058-3/19/$31.00 © 2019 IEEE

between conventional methodologies in terms of these:

- 92.99% detection for small sized sign boards;
- Enhanced feature learning with diverse background;
- Fast and efficient precautionary alert for driver;

Rest of paper is divided into six sections. Literature survey is discussed in section II, then section IV targets on Dataset details and description. Proposed methodology is listed in section III. Implementation details and results are discussed in section V.

II. Related Work

Since development of first autonomous car, traffic sign detection and recognition are essential parts of ADAS systems [4]. Before deep learning this issue was considered as traditional computer vision problem [12]. Then developments in deep learning helped researchers in achieving higher accuracies compared to conventional Computer vision & machine learning solutions [16]. Traditional image processing techniques [17], started to fade out with emergence of convolutional neural networks. Traffic sign detection systems contains two parts, first one is detection and other is recognition. In most of literature before rise of deep learning, detection is handled by using segmentation of image using specified thresholds. These thresholds were commonly defined as color space to obtain sign color from image. Then severity of that particular sign is categorized according to its color space [18]. Direct color matching was not adopted because of changing lightening conditions for sign boards on different time of day and varying environment. Maldonado Bascon, et al. [19] proposed a nice algorithm, based on support vector machines (SVMs) [20] for detection and recognition of road- signs automatically. They proposed a system which is built on SVM generalized properties. Their system takes three stages:

1) Color of pixel is used to segment out sign.
2) Usage of linear SVMs to classify shape of traffic-sign.
3) Using Gausian-Kernal SVMs for content recognition.

Their system performed fine in gaining low amount of false positive during recognition stage. Other than that, Mathias and Timofte, et. al [21] proposed set of diverse approaches. They used two difference datasets. German Traffic Sign Recognition benchmark and Belgium traffic sign dataset. Their proposed algorithm was based on detection through HOG features and sparse representation for classification. Pipeline of their system comprises of three stages:

- Feature extraction;
- Dimensionality reduction;
- Classification;

Results of this methodologies was 95%-99%. But it is also observed that most of proposed systems on these datasets also had accuracies in 95%-99% [21], which clearly depicts that vision-based systems performed very well on clean data. But problem arises when images captured with diverse background and size is used. All of listed accuracies and proposed systems started to fail due to arbitrary size of sign boards and backgrounds.

To overcome this issue deep neural networks had been adopted. In deep learning, there is a common practice of increasing number of layers in network to achieve higher accuracies for object detection. VGG-Net [22] and Google Net [17] are classical and popular examples of these kind of networks. These models are considered as highly complex in terms of parameters which heads up to high computation cost also. Alex Net [23] winning of ImageNet challenge in 2012 ILSVRC [24], helped a bit to reduce complexity level of Deep Neural Networks. Then development of state-of-the-art detection models Fast [25] and Faster R-CNN [14] stream lined production of good accuracies for object detection. However, it doesn't perform well on detection of small objects, which means significant drop in results for object detection.

Above discussion can be deduced into two 2 problems, one is bounding box size and other is poor appearance and very small structure objects. To elaborate first point of conclusion, consider PASCAL VOC wide Benchmark, in its each bounding box for target object is created by ratio of 20% of whole image. Which means it has more noisy features than required features of targeted object. For second point of conclusion, consider our problem statement for this paper: traffic signs board on roads usually occupy very small part of whole scene and most of the times its hard to clearly locate traffic sign board due to sign board small size. Resulting bad or ignored features of targeted object. By the fact that deeper neural networks perform better and increment in layers enhances learning ability, researchers did their best efforts [26] [26] to design another neural network architecture named as ResNet-101 [26]. This network comprises of 101 layers. Addition of residual modules and more convolutional layers to its network made it more efficient for abundant feature extraction of small sized objects.

Hereby problem of computational cost still remains, deeper network increases feature selection flexibility for model and also become uncontrollable while training. To reduce this computation cost we proposed solution by designing architecture based on Inception [27] in combination with RPN. Inception-v1 was introduced in Szegedy et al. 2015a [17] as GoogleLe Net. Inception architecture was then refined in later research. First refinement was made by introducing batch normalization [28] and named as inception-v2, second was adding additional factorization to it which is known as inception-v3 [29].

III. Proposed Methodology

Our proposed solution comprises of 2 main sections to detect and identify traffic signs correctly. First section extract feature map of input image by using modified inception-v3, second proposes regions for detected objects and classifies them.

A. Feature Extraction

Features extraction layers (FEL) are backbone of any deep learning network [14], which means that learning ability of a model strictly depends upon FEL.

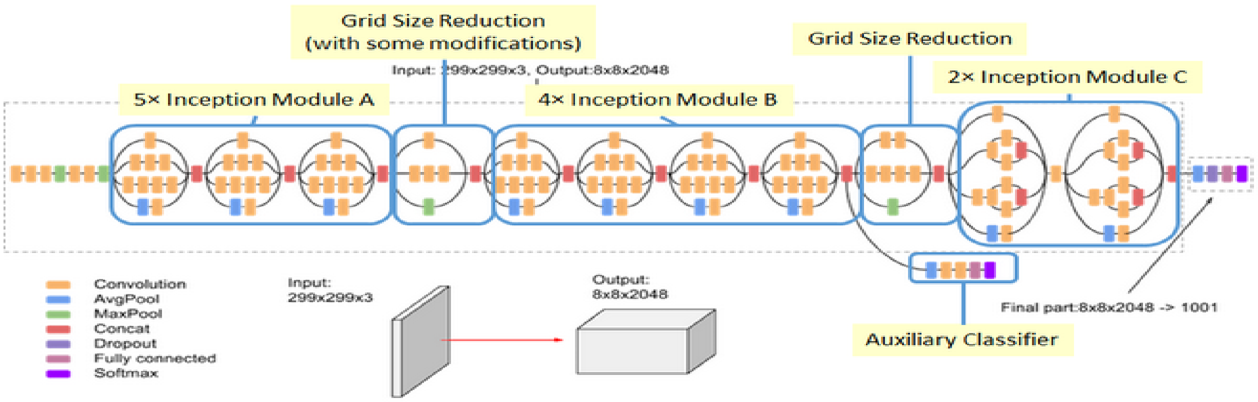

Fig. 1. Inception v3 model Architecture

That is why, our target is to extract accurate features of sign board in whole scene, so rest of the image features are not valuable for specified task. By using conventional methods [30] of features extraction, it has to be specified that which features to obtain, either global or local. Features are dependent on filter size, bigger filter-size utilizes global features and smaller filter size exploits local features. However, in case of variable object size, problem of selection between local and global features, turn into bottleneck issue. Solution to this problem was proposed with Inception network in fig. 1 [27].

Inception layers are essentially convolution layers, but they provide capability of variable filter sizes. We have exploited small part of inception-v3 architecture for obtaining variable feature map. Our input image passes from multiple convolution layers, illustrated in original inception-v3 fig 2, to calculate feature map. Moreover, the usage of feature maps obtained by inception-v3 are computationally inexpensive which helps to produce quick prediction [27].

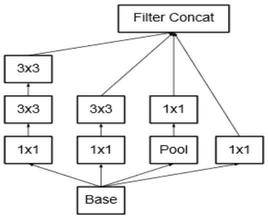

Fig. 2. Inception Module

To understand more consider fig 3, two [3x3] convolution filter replaced one 5x5 filter which eventually reduced number of parameters by 28%, mathematically this difference is shown below. [5x5] filter in 1 layer: [5x5] = 25 [3x3] filters in 2 layers: [3x3]+[3x3] = 18

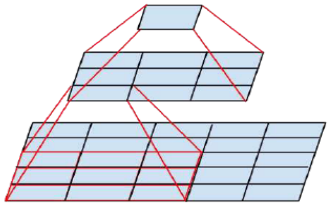

Fig. 3. Feature Extraction

In proposed methodology we used reduced version of inception-v3 by breaking original network at “3xInception” layer. Table 1 shows comparison network.

To extract features, we used [1025x1025x3] sized image to apply initial 3 convolutions, then 1 layer of pooling [3x3]. Finally, 3 convolution layers were applied before passing it to Inception module as mentioned in Table I. Resulting feature map of size [128x128x288] was used to pass in Region Proposed Network.

TABLE I
ARCHITECTURAL DETAIL OF INCEPTION

Type	**Patch Size/Stride**	**Original Inception V3 Input Size**	**Modified Inception V3 Input Size**
Conv	3x3/2	299x299x3	1025x1025x3
Conv	3x3/1	149x149x32	512x512x32
Conv Padded	3x3/1	147x147x32	510x510x32
Pool	3x3/2	147x147x64	510x510x64
Conv	3x3/1	73x73x64	254x254x64
Conv	3x3/2	71x71x80	252x252x80
Conv	3x3/1	35x35x192	128x128x192
3 x Inception	Fig 1.3(b)	35x35x288	128x128x288
5 x Inception	-	17x17x768	-
2 x Inception	-	8x8x1280	-
Pool	8x8	8x8x2048	-
Linear	Logits	1x1x2048	-
Softmax	Classifier	1x1x1000	-

B. Region Proposed Network

After extraction of feature map using inception it was passed into Region Proposed Netwrok (RPN). RPN layer

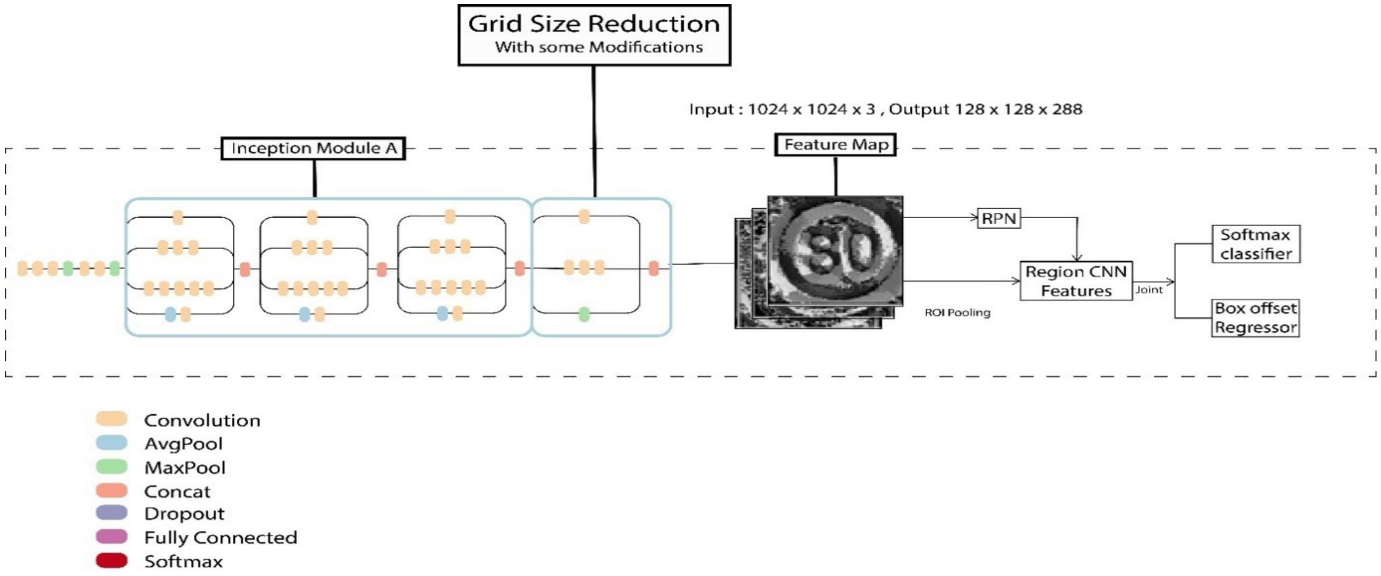

Fig. 4. Proposed Architecture

plots multiple bounding boxes with score bind to each box as result of detected object. These bounding boxes cannot describe or classify objects at this stage. But these boxes are eventually examined by classifier and regressor to identify object occurrence in given input image.

Basic idea of RPN is to classify Background and foreground in input image [14]. Like in out case we want anchor to drop on traffic sign board in whole scene. It is addressed in training by providing groundtruth boxes in training data. Target of RPN is to learn foreground by identifying anchors which has maximum overlapping with groundtruth. Other than that learning lesser overlapping anchors as background.

C. ROI Pooling & Classification Layer

RPN layers propose different sized regions. Which means CNN feature map of different sizes. It was not easy to construct efficient structure for different feature sizes. To handle this limitation Region of Interest (ROI) pooling [14] is applied, which reduces feature map into identical size. ROI pooling works little different from Max-Pooling, it does not contain fix size. ROI pooling divides input feature map in fixed number which are roughly equal, after that Max-pooling is applied to each of it. Therefore, ROI pooling finally generates output of same size as division applied to its feature map (fig. 5).

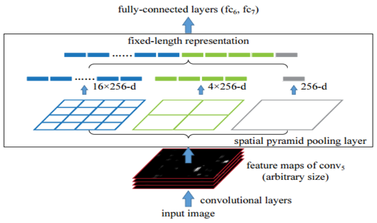
fully-connected layers (fc_6, fc_7)
fixed-length representation
16×256-d
4×256-d
256-d
spatial pyramid pooling layer
feature maps of $conv_5$
(arbitrary size)
convolutional layers
input image

Fig. 5. ROI Pooling

Output of fixed Region of Interest pooling is then used as input of classification layer. ROI pooling output produces good input for classifier and regressor to choose best prediction. Finally, SoftMax is used in dense layer to normalize final output Fig. 4.

$$\text{Softmax score: } s_i = \frac{e^{s_i}}{\sum_j e^{s_j}} \forall\ i\ in\ \{1, 2..C\} \tag{1}$$

IV. Dataset

In order to train DNN with varying size traffic sign boards, Chinese researchers [12] generated new dataset by installing cameras on city roads and highways of china (fig. 6). They named it Tsinghua-Tencent 100k Traffic Signs dataset. To train our network we used this dataset. This dataset is majorly divided into 3 classes. These classes contain unique color with alphabetic representation Fig 7.

Fig. 6. Tsinghua-Tencent 100k Traffic Signs-Dataset Images

Yellow Fig 7 colored boards are used for representing Warnings and assigned with letter ‘w’, Red Fig 7 represents for Prohibitory with ‘p’ and blue Fig 7 is used to describe Mandatory with ‘i’. These classes are then divided into sub classes. Given Table II describe more about it.

TABLE II
DATASET CLASSES REPRESENTATION

Class	Color	Representation Letter	Sub Classes
1	Yellow	w	67
2	Red	p	28+15 Special Classes
3	Blue	i	15+2 Special Classes

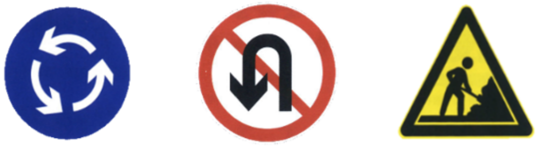

Fig. 7. Tsinghua-Tencent 100k dataset Classes

There are total 9,180 annotated images containing traffic signs with different location, size and distance. but there are another 7,643 images without traffic signs to increase robustness of data and add generalization while training. All of this dataset is recorded with panoramic view using 6 SLR cameras [12]. Then joined together to form better and wider capture area. This dataset was collected on 300 Chinese cities and linked road networks.

As our dataset image original size [2048x2048x3], which is very expensive for training, we resized all dataset images to [1024x1024x3]. This input size of image is still very big to create batch and load in to memory, so we generated .npy files for whole dataset to reduce data loading time. Which helped in training our model with batch size of 32.

V. EXPERIMENTS

A. Implementation Details

We used Nvidia GPU 1080Ti with 11Gb memory, machine used for training was i7-7th generation with 16Gb RAM. Our model is developed using Keras backend. Hyper parameters for trainings are discussed further. Learning rate is set to 0.0001 with decay rate of 0.7. Momentum is set to 0.8. initially Adam is used to boost learning ability of network from scratch then after 20 epochs we changed it to SGD. Mean squared error is used to calculate Loss for learning weights.

$$\text{Mean Square Error: } \frac{1}{i}\sum_{j=1}^{i}(\hat{X}_j - X_j)^2 \qquad (2)$$

B. Results

The model architecture was built using TensorFlow's api named Keras. Training of model is performed on NVIDIA 1080 TI GPU having 11 GB memory and 3584 CUDA cores. Mean Square Error (MSE) is used for loss calculation and SGD as optimizer with dynamic learning rate described in table 4. Model is trained for 300 epochs with dropout ratio of 0.4 to avoid model from over fitting as shown in table III. We found 0.4 is best dropout rate which results training and validation accuracy very close. Training and test accuracy with different dropout ratio are shown in table IV.

TABLE III
LEARNING RATE AGAINST EPOCH

Epochs	Learning Rate
1-50	0.001
51-150	0.0001
151-250	0.00001

TABLE IV
TRAINING/TESTING ACCURACY ON DIFFERENT DROPOUT RATIO

Dropout Ratio	Training Accuracy	Validation Accuracy
0.1	98.65%	68.95%
0.3	96.85%	82.68%
0.4	**94.18%**	**92.99%**
0.5	86.95%	85.26%

We used 80% data to train out model and remaining 20% is used to evaluate the learning of model. Loss graph for training and validation is shown in fig. 8. To evaluate the model, we split the dataset into three main category small, medium and large traffic signs.

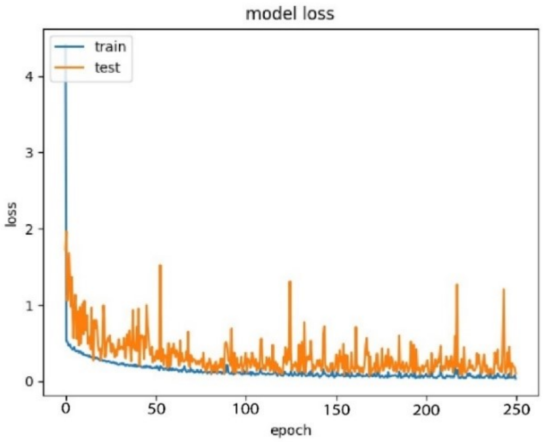

Fig. 8. Training/Testing Loss

We compared accuracy of our model with Faster R-CNN with VGG16 and Zhe Zhu [12] et al. Comparison results are shown in Fig. 9. As shown in fig. 8 Faster R-CNN performs good for large size of traffic signs than Zhe Zhu [12] and bad for small traffic signs and on other hand Zhe Zhu [12] performs better for small traffic signs than faster RCNN. Our approach outperforms in both scenarios.

VI. CONCLUSION

Increase of vehicles in modern era and growing demand of autonomous vehicles made it compulsory to adopt Advanced driving assistance systems. To improve automatic sign recognition for these systems we developed a new state of the art deep learning model architecture for small and large size traffic sign detection and classification. Previous models either performs good for large size traffic signs or small size traffic signs. The results show our model out performed in both cases by detecting varying sign boards sizes. In future,

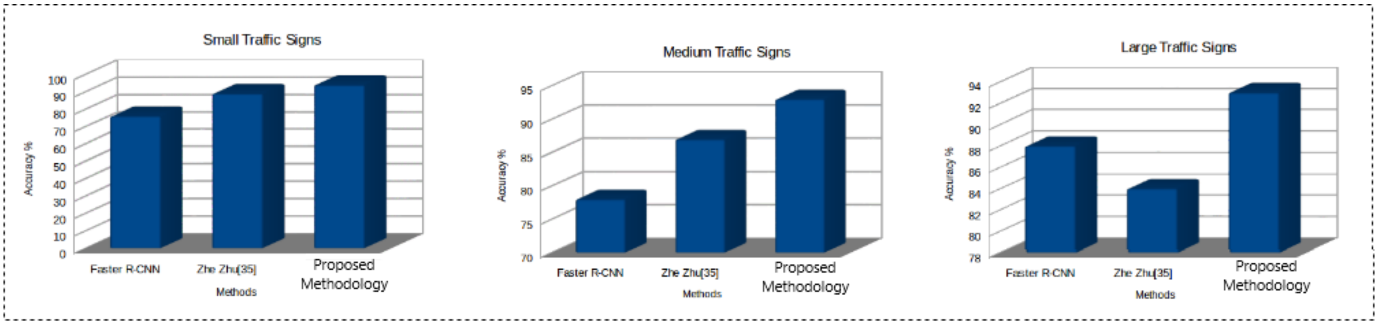

Fig. 9. Comparative Analysis

we are going to work on the response time of this algorithm by optimizing its parameters. We also plan to develop lighter version of proposed model. So, it can work on mobile devices as well as Embedded systems.

References

[1] WHO, "Road traffic injuries," https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries, December 2018, accessed on 19-04-2019.

[2] M. Lu, K. Wevers, and R. Van Der Heijden, "Technical feasibility of advanced driver assistance systems (adas) for road traffic safety," *Transportation Planning and Technology*, vol. 28, no. 3, pp. 167–187, 2005.

[3] infoPlus, "Modern car navigation systems and their features," https://www.infopulse.com/blog/modern-car-navigation-systems-and-their-features/, April 2018.

[4] R. Viereckl, D. Ahlemann, A. Koster, and S. Jursch, "Connected car study 2015: Racing ahead with autonomous cars and digital innovation," *Strategy& http://www. strategyand. pwc. com/reports/connected-car-2015-study [Last accessed 7 March 2016]*, 2015.

[5] M. J. Chiappetta, "Celestial navigation system for an autonomous vehicle," Mar. 3 2015, uS Patent 8,972,052.

[6] B. Philanthropies, "Road deaths and injuries hold back economic growth in developing countries," https://www.worldbank.org/en/news/press-release/2018/01/09/road-deaths-and-injuries-hold-back-economic-growth-in-developing-countries, January 2018, accessed on 19-04-2019.

[7] theroadtochangeindia, "Road signs – an important traffic management tool." https://www.worldbank.org/en/news/press-release/2018/01/09/road-deaths-and-injuries-hold-back-economic-growth-in-developing-countries, November 2018, accessed on 19-04-2019.

[8] Z. Huang, Y. Yu, J. Gu, and H. Liu, "An efficient method for traffic sign recognition based on extreme learning machine," *IEEE transactions on cybernetics*, vol. 47, no. 4, pp. 920–933, 2016.

[9] Y.-L. Chen, B.-F. Wu, H.-Y. Huang, and C.-J. Fan, "A real-time vision system for nighttime vehicle detection and traffic surveillance," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 5, pp. 2030–2044, 2010.

[10] Y. Xu, J. Dong, B. Zhang, and D. Xu, "Background modeling methods in video analysis: A review and comparative evaluation," *CAAI Transactions on Intelligence Technology*, vol. 1, no. 1, pp. 43–60, 2016.

[11] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[12] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2110–2118.

[13] Y. Jeon and J. Kim, "Active convolution: Learning the shape of convolution for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4201–4209.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[15] G. Yannis, E. Papadimitriou, and K. Folla, "Effect of gdp changes on road traffic fatalities," *Safety science*, vol. 63, pp. 42–49, 2014.

[16] L. Zhang, J. Tan, D. Han, and H. Zhu, "From machine learning to deep learning: progress in machine intelligence for rational drug discovery," *Drug discovery today*, vol. 22, no. 11, pp. 1680–1685, 2017.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[18] L. Liu, X. Wang, and P. Pan, "Traffic sign detecting method and traffic sign detecting device," Mar. 17 2015, uS Patent 8,983,136.

[19] V. A. Prisacariu, R. Timofte, K. Zimmermann, I. Reid, and L. Van Gool, "Integrating object detection with 3d tracking towards a better driver assistance system," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3344–3347.

[20] N. Guenther and M. Schonlau, "Support vector machines," *The Stata Journal*, vol. 16, no. 4, pp. 917–937, 2016.

[21] S. Maldonado-Bascón, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gómez-Moreno, and F. López-Ferreras, "Road-sign detection and recognition based on support vector machines," *IEEE transactions on intelligent transportation systems*, vol. 8, no. 2, pp. 264–278, 2007.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[25] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[30] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4544–4554, 2016.