Large Scale Traffic Surveillance : Vehicle Detection and classification using Cascade Classifier and Convolutional Neural Network

Shaif Chowdhury
Dept. of Information Technology
Institute of Engineering & Management, Kolkata
chowdhuryshaif@rediffmail.com

Abstract — In this Paper, we are presenting a traffic surveillance system for detection and classification of vehicles in large scale videos. Vehicle detection is crucial part of Road safety. There are lots of different intelligent systems proposed for traffic surveillance. The system presented here is based on two steps, a descriptor of the image type haar-like, and a classifier type convolutional neural networks. A cascade classifier is used to extract objects rapidly and a neural network is used for final classification of cars. In case of Haar Cascades, the learning of the system is performed on a set of positive images (vehicles) and negative images (non-vehicle), and the test is done on another set of scenes. For the second, we have used faster R-CNN architecture. The cascade classifier gives faster processing time and Neural Network is used to increase the detection rate.

Keywords — Surveillance, CCTV, Haar Cascades, R-CNN, Neural Network, Deep Learning, caffe

I. *Introduction*

No of motor vehicles have increased tremendously in developed countries. That coupled with the advancement of CCTV cameras has led to an increased demand for automated traffic surveillance systems. Intelligent traffic monitoring system can improve road safety and decrease criminal activities too. One of the most important part of traffic surveillance is vehicle recognition. An intelligent system capable of vehicle recognition can also be used in driving assistance. The vehicle detection is essential in intelligent systems as it can detect potentially dangerous situations with vehicles in advance to warn the driver.

In last few years there has been a tremendous increase in use of CCTV cameras for traffic regulation. Traffic cameras are innovative and extremely useful video surveillance. Traffic cameras are used in various ways from monitoring traffic, issuing tickets for moving violations to record traffic patterns for future study.

In some countries like China, Japan, Singapore video surveillance market is forecasted to grow at more than 10% per year. The major driver here is the growing popularity of deep learning based techniques. There is a huge scope for deep learning-based video analytics servers and video management platforms with integrated video analytics. There are some notable startups already using deep learning for surveillance. For example, Camio which offers an app that lets a smartphone or tablet act as a surveillance camera. It captures videos through phone camera and uses machine learning to point out most of the significant events captured by a user's camera. Camio is also expanding its use of artificial neural

networks to enable users to search for objects that are difficult to identify like cats, dogs, bikes, trucks, and packages.

I-Corps is another company based on Real-Time Traffic Congestion Detection. Detecting traffic congestion involves two components: (a) vehicle detection and tracking; and (b) event classification. Computer vision is used for detecting vehicles . By analyzing consecutive frames, vehicle speed and relative locations will be estimated. For the classification of events, the team will use their expertise in graph kernels and machine learning. A sequence of consecutive frames can be used to create a graph, where the nodes represent vehicles labeled with local features, such as speed and location. Neighboring nodes in the graph are connected by edges labeled with the distance between their respective vehicles. A fast kernel function for graphs will be developed and used by a binary classifier which will be trained for the task of recognizing traffic congestion. As an extension, a multi-class classifier can also be trained to distinguish different types of traffic events, such as, high, moderate, or low traffic congestion, accident, or normal traffic.

Data is the lifeblood of the modern world. Today, it’s being captured by more than 500 million cameras worldwide, and that number is growing exponentially. This is creating a tsunami of data that is just impossible for human eye to analyze. Artificial Intelligence is the key to turning this information into insight use it to capture, inspect, and analyze data to impact everything from public safety, traffic, and parking management to law enforcement and city services.

This paper is a contribution in the field of intelligent systems capable of detecting vehicles. Here we present a vehicle detection system combining two algorithms, the first is an image processing algorithm and the second is an algorithm of artificial intelligence. The image processing algorithm aims to detect vehicles using haar classifier, and the algorithm of artificial intelligence uses convolutional neural networks to classify these vehicles.

Neural network is an information processing system patterned according to the operation of neurons in the human brain. For object detection neural network is used in similar way to optic nerves in human visual processing. Neural network usually involves thousands, or even millions of artificial neurons called units operating in parallel and arranged in tiers. The first tiers are input units designed to receives the raw input information. The successive tiers receive inputs from the previous layer - similar to the way neurons further from the optic nerve receive signals from those closer to it. The last tier responds using the information it's learned.

Neural networks are known to be for adaptive, which means they modify themselves as they learn from initial training and moves to subsequent units. The most basic learning model is centered on weighting the input streams. This design called a feedforward network. Each subsequent input is multiplied by the weights of the connections they travel along. All units add up the inputs received in this way and if the sum is more than a certain threshold value, the next node is triggered.

There's also an element of feedback involved called backpropagation. This technique makes use of hidden nodes. The output of a network is compared with the output it was meant to produce, and the difference between them is used to modify the weights of the connections between the units in the network, working from the output units through the hidden units to the input units—going backward.

ConvNet makes one assumption that is the input is a multi-channeled image, which allows it to encode certain properties into the architecture. unlike a regular Neural Network, the layers of a ConvNet have neurons arranged in 3 dimensions: width, height, depth. Other than that it's made up of neurons with learnable weights and biases. Each neuron

receives several inputs, takes a weighted sum over them, pass it through an activation function and responds with an output. The whole network has a loss function and everything else about neural networks still apply on CNNs.

The paper is divided as follows. The following section presents previous research on the detection of the vehicle. Section 3 describes in detail the descriptor and the classifier along with results. The last section is devoted to the conclusion.

II. Related Work

Until now, many researchers have proposed different vehicle detection systems. A simple solution to the problem is the exhaustive search of all possible positions in the processed image. But this solution is unsatisfactory due to low calculation speed3. To solve the problem many researchers have proposed learning algorithm to extract the features from the objects to be detected.

The most used system here is proposed by Viola and Jones [1] (Viola and Jones, n.d.) , which consists in extracting the haar like features [2] (Oren et al., n.d.) , [3] (Destrero, Odone, and Verri 2007) and to cascade a set of weak classifiers to construct a strong classifier. Here [4] (H. Wang and Zhang 2014) , the authors used a 2 step process. The first step finds the shadow areas of the vehicle using Haar algorithm and as AdaBoost classifier. The second step is done by applying the combined treatment of regions of interest with HOG and SVM algorithm as the classifier to verify the presence of the vehicle and then use the K-means algorithm to increase the detection rate. In [5] (Sivaraman and Trivedi 2010) , the authors present a framework for active learning for robust recognition of road vehicles, using the supervised learning algorithm applied wit

a particle filter.

Deep learning is currently the most successful research direction in the field of machine learning. Since proposed in 2006 [6] (Hinton 2006) , it has made huge progress in areas of information processing such as voice, text, image, video and

so on.Deep learning is combination of a group of machine learning techniques that can learn features hierarchically from lower level to higher level by building a deep architecture. It has the ability to automatically learn features at multiple levels, which makes the system be able to learn complex mapping function $f: X \rightarrow Y$ directly from data, without help of the human-crafted features. In recent years, deep learning has improved the performance in computer vision by a huge margin. And in image classification, object localization, scene classification, object detection, almost all the top algorithms are based on deep learning. Here we would discuss the progress of deep learning in object detection.

Deep learning tracker (DLT) [7] (Bebis et al. 2016) is the first deep network for tracking task, in which the idea "off-line pretrain+online fine-tune" is proposed. Convolutional Neural Network : Due to the use in image recognition, CNN has become the mainstream deep model in computer vision and in object tracking. The conceptuses off-line training of large-scale CNN as both classifier and tracker. From DLT onwards, a large number of CNN-based tracking algorithms appeared, and two representative of them are fully convolutional network tracker [8] (L. Wang et al. 2015) (FCNT) and multi-domain convolutional neural network [9] (Nam and Han 2016) . Deep convolutional neural networks have been improved substantially to use in image classification and other recognition tasks. Since their introduction in the early 1990s , convolutional neural networks have consistently been competitive with other techniques for image classification and recognition. Recently, they have pulled away from competing methods due the availability of larger datasets, better models and training algorithms and the

availability of GPU computing to enable investigation of larger and deeper models.

III. Proposed System

This section is devoted to the description of the proposed vehicle detection system. This system is composed of two parts: learning section and detection section. The learning section is done in two phases: Extraction phase of the features for learning: During this phase a descriptor based on haar wavelet is extracted for each image in the training set, a feature vector. This database contains a set of positive images fig.1.a (with car) and negative fig.1.b (no vehicle).

Classification phase: The car images from the first phase are fed as an input to the neural network. The architecture of the selected neural network is Faster R-CNN which has been proven to be robust and fast.

Haar-like features :

Haar feature technique was initially proposed as a face detection technique by voila and jones. It's a machine learning based technique where a classifier is trained with a few hundred positive images and a few hundred negative images. Once the classifier is trained, it is applied on a region of interest. If the object is present then a positive output is given and negative otherwise.

In cascade classifier the original image is partitioned in rectangular patches, each of which is passed through different stages and classified as positive or negative. In case of Haar cascades local features are calculated by subtracting the sum of a subregion of the feature from the sum of the remaining region of the feature. The image here shows an extended set of twisted(45 deg) Haar-like feature :

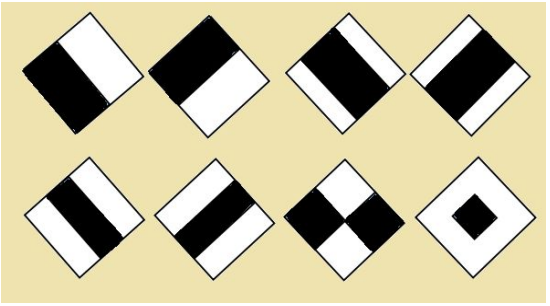


Fig 1.

For calculating haar features Viola and Jones introduced the concept of the integral image eq.1, which gives a representation of an input image and reduces the computation time of these features. The speed of the integral image in the calculation shows a rectangular sum using only four references, then the difference between two adjacent rectangles, can be calculated with only six references and eight for three rectangles.

$$ii(x,y) = \sum_{x \le x', y \le y'} i(x',y') \qquad \text{eq.1}$$

Where,

$ii(x,y)$ *is the integral image.*

$i(x,y)$ *is original image.*

This allows calculation of Haar windows at various scales in an image.

$$s(x,y) = s(x,y-1) + i(x,y)$$

$$ii(x,y) = ii(x-1,y) + s(x,y)$$

This approach uses only 2 rectangular filters (horizontal and vertical) at different scales: 2x2, 4x4, 8x8 and 16x16.

Neural Network :

As we described above, a simple ConvNet is a sequence of layers, and every layer of a ConvNet transforms one volume of activations to

another through a differentiable function. The main types of layers used to build ConvNet are: Convolutional Layer, Pooling Layer, and Fully-Connected Layer. We will stack these layers to form a full Here, we'll look at R-CNN (Regional CNN) and its descendants Fast R-CNN, and Faster R-CNN.

But how do we find out where these bounding boxes are? R-CNN would propose a bunch of boxes in the image to see if any of them actually correspond to an object.

R-CNN creates region proposals, using a technique called Selective Search which you can read about here. At a high level, Selective Search looks at the image through windows of different sizes, and for each size tries to group together adjacent pixels by texture, color, or intensity to identify objects.

R-CNN works really well, but is quite slow. It requires a forward pass of the CNN for every single region proposal for every single image.

It has to train three different models separately - the CNN to generate image features, the classifier that predicts the class, and the regression model to tighten the bounding boxes. This makes the pipeline extremely hard to train.

Faster R-CNN has two networks: region proposal network (RPN) for generating region proposals and a network using these proposals to detect objects. The main different here with Fast R-CNN is that the later uses selective search to generate region proposals. The time taken to generate region proposals is much smaller in case of RPN than selective search, when RPN shares the most computation with the object detection network. Briefly, RPN ranks region boxes (called anchors) and proposes the ones most likely containing objects.

Faster R-CNN is an extension of the R-CNN and Fast R-CNN object detection techniques. The difference between them is how they select regions to process and how those regions are classified. R-CNN and Fast R-CNN use a region proposal algorithm as a pre-processing step before running the CNN. The proposal algorithms are generally techniques such as EdgeBoxes or Selective Search, which are independent of the CNN. In the case of Fast R-CNN, the use of these techniques becomes the processing bottleneck compared to running the CNN. Faster R-CNN solves this issue by implementing the region proposal mechanism using the CNN and thereby making region proposal a part of the CNN training and prediction steps.

The image processing algorithm aims to extract the features of a vehicle using the descriptor haar, and the algorithm of artificial intelligence uses artificial neural networks to classify and detect these vehicles.

The detection step consists in cutting the original image into zones and then extract the feature vector of each zone with the same algorithm used during the learning phase. Finally using the same classifier will be used to verify the presence of the type of vehicle in the treated area image.

Here is the logic flow of CNN based system

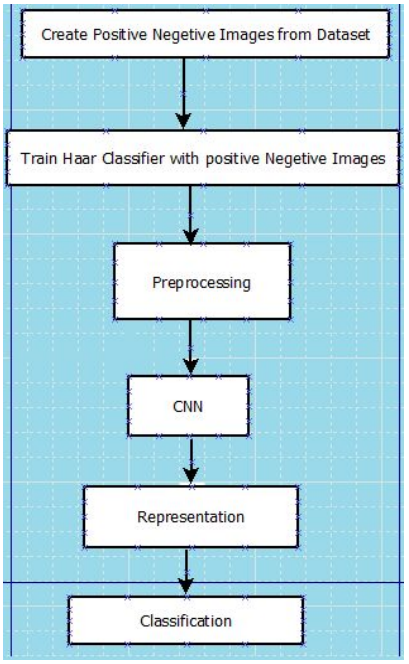


Fig 2.

IV. Results

The database of our detection system is composed of a set of positive images: the positive database contains 500 examples. The examples here consist of different cars from views.

Fig 3.

A set of negative that contain randomly selected negative patches. The test database: The video used for testing was originally filmed at Sherbrooke/Amherst intersection in Montreal [10] (Nam and Han 2016; Jodoin, Bilodeau, and Saunier 2014) . The video resolution is 800x600. For the evaluation, a 1001 frames (30 fps) part of the video was chosen with 15 cars and 5 pedestrians. We have cropped the images according to need.

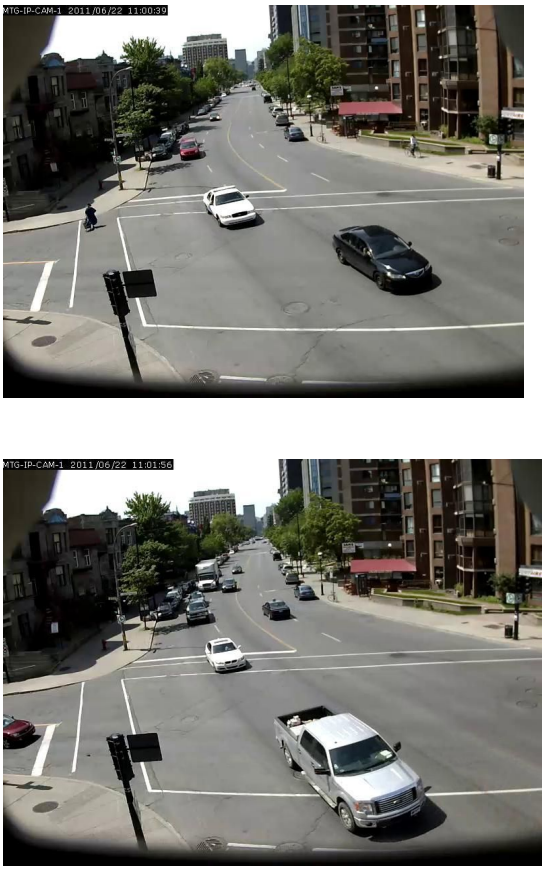


Fig 4.

The neural network is implemented using caffe on a PC-type platform for features: (Intel® Core™ i3 CPU (2,40GHZ), RAM 4,00 Go, Linux Ubuntu 14.0):

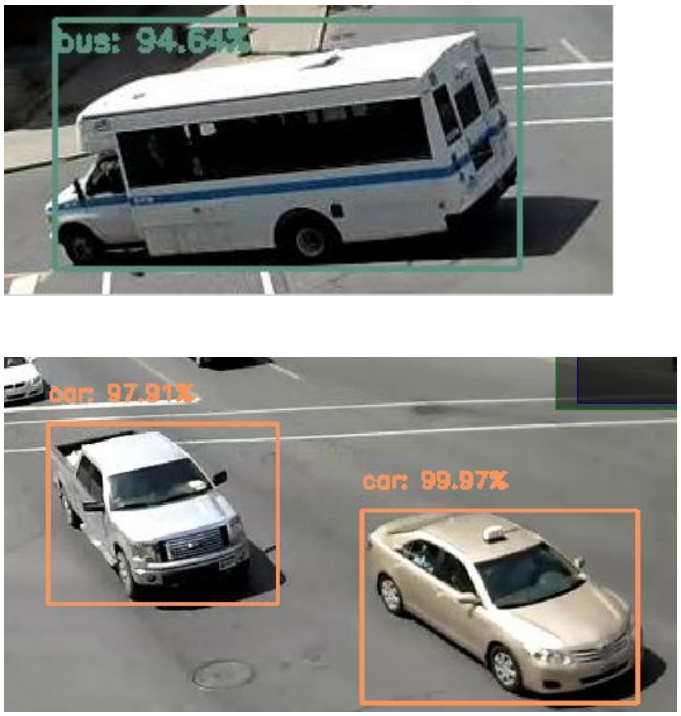
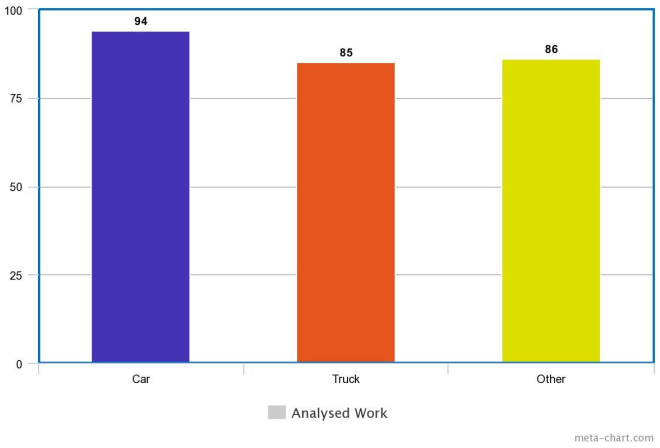


Fig 5.

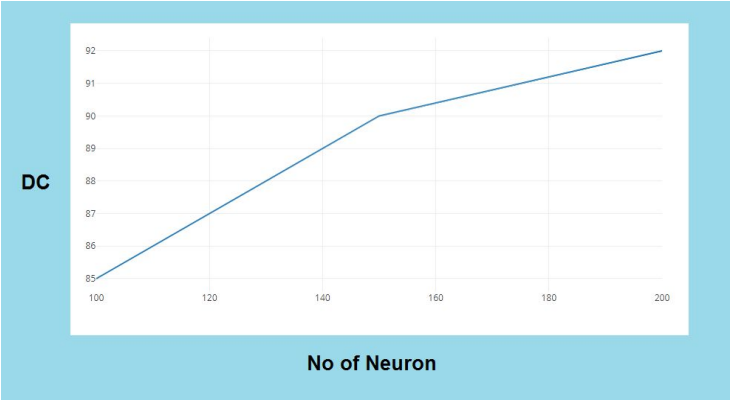
The detection produces some false positive results. The overall error rate reduces with increasing number of positive images. So, more appropriate data along with some hard-negative patches can be used to fine tune the model.

Analysed Work	Number of Samples	TRUE	False	Percentage
Car	500	470	30	0.94
Truck	100	85	15	0.85
Other	80	69	11	0.86



The experimental results show the detector performance with a high detection rate reached more than 90% at 150+ neurons. The results also shows that the execution time per image naturally increases with the number of neuron, but this increase does not affect the speed of this detection system.

No of neurons	DC	Execution Time
100	85	0.43
150	90.3	0.56
200	92.1	0.69



The average classification accuracy here is 92% along with a low standard deviation of classification error.

V.*Conclusion*

The proposed system has combined efficient algorithms in of image processing and in artificial intelligence.This study shows that the combination of haar like classifier and CNN is a good candidate for object detection. The system is presented with highway traffic in mind. So, a future research would be to test it in real time system using a raspberry PI and a digital camera. That would present several scope of improvements to make it usable in real applications.

References

1. Viola, P., and M. Jones. n.d. “Rapid Object Detection Using a Boosted Cascade of Simple Features.” In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. https://doi.org/10.1109/cvpr.2001.990517.
2. Oren, M., C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. n.d. “Pedestrian Detection Using Wavelet Templates.” In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.1997.609319.
3. Destrero, Augusto, Francesca Odone, and Alessandro Verri. 2007. “A Trainable System for Face Detection in Unconstrained Environments.” In *14th International Conference on Image Analysis and Processing (ICIAP 2007)*. https://doi.org/10.1109/iciap.2007.4362812.
4. Wang, Huan, and Haichuan Zhang. 2014. “A Hybrid Method of Vehicle Detection Based on Computer Vision for Intelligent Transportation System.” *International Journal of Multimedia and Ubiquitous Engineering* 9 (6):105–18.
5. Sivaraman, Sayanan, and Mohan Manubhai Trivedi. 2010. “A General Active-Learning Framework for On-Road Vehicle Recognition and Tracking.” *IEEE Transactions on Intelligent Transportation Systems* 11 (2):267–76.
6. Hinton, G. E. 2006. “Reducing the Dimensionality of Data with Neural Networks.” *Science* 313 (5786):504–7.
7. Bebis, George, Richard Boyle, Bahram Parvin, Darko Koracin, Fatih Porikli, Sandra Skaff, Alireza Entezari, et al. 2016. *Advances in Visual Computing: 12th International Symposium, ISVC 2016, Las Vegas, NV, USA, December 12-14, 2016, Proceedings*. Springer.
8. Wang, Lijun, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. 2015. “Visual Tracking with Fully Convolutional Networks.” In *2015 IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2015.357.
9. Nam, Hyeonseob, and Bohyung Han. 2016. “Learning Multi-Domain Convolutional Neural Networks for Visual Tracking.” In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.465.
10. Jodoin, Jean-Philippe, Guillaume-Alexandre Bilodeau, and Nicolas Saunier. 2014. “Urban Tracker: Multiple Object Tracking in Urban Mixed Traffic.” In *IEEE Winter Conference on Applications of Computer Vision*. https://doi.org/10.1109/wacv.2014.6836010.