Traffic Signs Recognition and Distance Estimation using a Monocular Camera

Shadi Saleh
M.Sc., Chair of Computer Engineering, Chemnitz University of Technology
Chemnitz, Germany D-09111
shadi.saleh@informatik.tu-chemnitz.de

Sinan A. Khwandah
Dr, Warsash School of Maritime Science &Technology, Southampton Solent University
Southampton, United Kingdom
Sinan.khwandah@solent.ac.uk

Ans Mumtaz
M.Sc., Chair of Computer Engineering, Chemnitz University of Technology
Chemnitz, Germany D-09111
ans.mumtaz@s2015.tu-chemnitz.de

Ariane Heller
Dr, Chair of Computer Engineering, Chemnitz University of Technology
Chemnitz, Germany D-09111
ariane.heller@informatik.tu-chemnitz.de

Wolfram Hardt
Prof, Chair of Computer Engineering, Chemnitz University of Technology
Chemnitz, Germany D-09111
hardt@cs.tu-chemnitz.de

Abstract: Traffic signs are integral component of road transport infrastructure. They deliver essential driving information to road users, which in turn require them to adapt their driving behavior to road regulations. In this study, the deep learning model is implemented with You Only Look Once (YOLO) based on active learning approach in order to minimize the size of the labeled dataset and provide higher accuracy. This is an efficient approach using a single monocular camera to perform real-time traffic signs recognition and distance estimation between traffic sign and camera pose. YOLO is one of the faster object detection algorithms, and it is a very good choice when real-time detection is needed, without loss of too much accuracy. The active learning algorithm will bridge the gap between labeled and unlabeled data, thus, only queries the samples that would lead to increase the accuracy. The aim of this work is to alarm and notify the drivers without having them to switch their focus. The results of the performed experiments show that about 97% recognition accuracies could be achieved with real-time capability in different real-world scenarios.

Keywords: Deep learning, YOLO, Active learning, Distance estimation, Monocular camera, Traffic signs recognition.

1 Introduction

Millions of people are seriously injured in vehicle accidents every year. Usually road accidents are caused by negligence, ignorance of the rules and traffic signs. All signals contribute to keep order in road traffic and are also intended to reduce the number and severity of road accidents. The traffic signs are placed in predefined areas to ensure the driver's safety. Nowadays, the recognition of traffic signs has been the subject of interest for automotive community and is even a very valuable characteristic of autonomous vehicles. The objective of automatic traffic sign recognition systems is to specify the locations and sizes of traffic signs in natural scene stream images (detection task) and then classify the detected traffic signs into their specific classes (classification task). The development of autonomous traffic sign recognition systems assists the driver in a variety of ways to ensure his safety, which includes the safety of other drivers and pedestrians. The main purpose of these systems is the recognition and identification of traffic signs while driving. Through these functions (detection and classification tasks), the systems can control and alert the driver to avoid hazards. Traffic signs contain important useful information that the driver may ignore due to driving fatigue or searching for address needs. In addition, drivers usually pay less attention to traffic signs in rough weather. Therefore, making efficiency initiatives, such as increasing driving safety and improving the automatic identification and road sign recognition system, is essential to help reducing road casualties. These improvements, although having a positive impact, face several external non-technical challenges, such as light variations, size and weather conditions, and signs in destroyed conditions, which can ultimately affect the performance of traffic sign recognition systems. Usually, Traffic signs vary slightly among countries and they are diverse and might include a wide range of colours, shapes, and dimensions. The high interest variability and dynamic occlusions can significantly challenge in the development of an automated method. Other challenges are uncontrolled illumination, resolution, image quality and occlusion and confusion with a man-made object. Figure1 shows some examples of non-ideal invalid and challenging traffic signs.

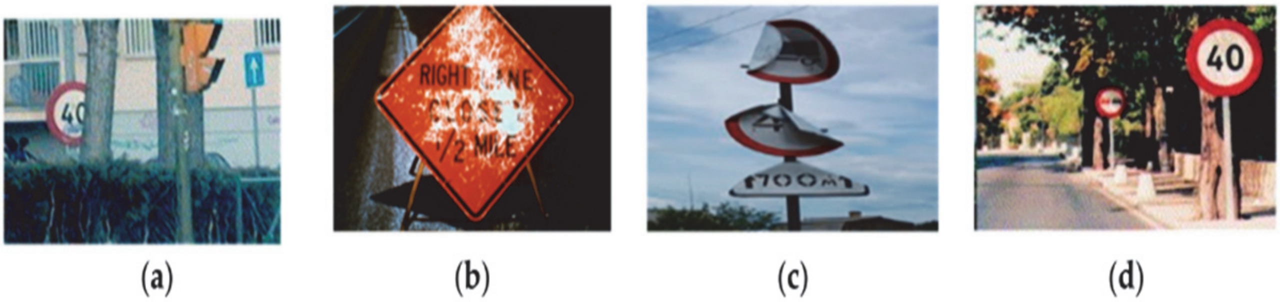


Figure 1 – Shows non-identical traffic signs, (a) Partial occlusion, blurred traffic sign, (c) destroyed traffic sign, (d) multiple traffic signs displayed at the same time.

The main concern with traffic sign recognition system is not how to recognize or identify a traffic sign with a high reminder in a still image. In fact, it is how to achieve high accuracy in high-resolution live video streaming with real-time capability [2][3]. In order to demonstrate the problem of false detection, a traffic sign system at least with 30 frames per second (108,000 frames per 1h video) was considered. Under the assumption that the system has a false positive accuracy of 1%, this means 17 false alarms were detected every minute (1071 in 1 hour) and 1 really positive and 68 false alarms within 4 minutes. Traffic signs have distinctive features which can be classified into separate sub-categories.. These are divided into five main categories according to shape and colour: Warning signs (red trianle), prohibition signs (red round), reservation signs (rectangular blue), mandatory signs (circular blue) and temporary signs (yellow triangle). In this study, only the speed signs and the stop sign are considered as use case studies in order to evaluate our approach. In the proposed approach, the system will supply the vehicle driver with real-time traffic sign information, the estimated distance between the monocular camera mounted on the vehicle, and the detected traffic sign. Then the system will notify the vehicle driver if the speed of the vehicle does not match the recognized traffic sign.

2 Problem Statement

Nowadays, the majority of studies are based on supervised learning to tackle the problem of traffic sign recognition. Furthermore, the performance of current solutions depends on large amounts on the labeled dataset for training and is limited to a predetermined labeled dataset, which is actually tricky and costly. Furthermore, they are not suitable for real-time applications. Some existing systems have identified the recognition and automated identification of traffic signs as a difficult problem in several important application areas, including updated driver control systems, road inspections and autonomous vehicles. Despite the fact that a lot of researches are carried out both in the field of identification and automatic recognition of symbol-based traffic signs and in the identification of texts in real scenarios. The speed and distance estimation systems were not implemented in the existing traffic sign recognition systems.

In this study, the deep learning model is implemented with You Only Look Once (YOLO) [4] based on an active learning algorithm in order to reduce the amount of labelled training dataset provide higher accuracy with minimum of labeled dataset. A monocular camera and deep learning model are used to perform real-time traffic sign detection and distance estimation between traffic signs and camera position. Where YOLO is one of the faster object detection algorithms, and it is a very good choice when real-time detection is required, without loss of too much accuracy. Active learning algorithm bridging the gap between labelled and unlabeled data, thus, only queries the most useful examples that would lead to increase the accuracy. The aim of this work is to alarm and notify the drivers without the objective of this work is to raise alarms and inform drivers without them having to change focus. Experimental results showed an accuracy of approximately 97% with real-time ability in different real scenarios.

3 Related Work

There has been a great interest in self-driving cars and Advanced Driver Assistance Systems (ADAS) for several years. In order for these systems to become more autonomous and driver assistance systems, the integration of traffic sign recognition systems is an essential requirement. Traffic sign recognition systems consist of two main components: classification and recognition. The classification component is strictly focused on the classification of the type of traffic sign after the traffic sign has been detected. The recognition component, on the other hand, will concentrate on locating the traffic sign in a sequence of images. These systems can be classified into two main categories, one is based on traditional object detection algorithms which combined the properties of traffic signs such as color-based and shape-based. The other one is identified by learning-based methods, such as machine learning approaches and deep learning techniques, which have the ability to self-learn various features. The goal is to demonstrate a light structure of some of the current and effective approaches to traffic sign recognition and classification. In the following are brief overviews of the state-of-the-art traffic sign recognition researches.

3.1.1 Traditional Object Detection Techniques

3.1.1.1 Colour Spaces

Traditional recognition techniques mainly relied on feature extraction methods. Generally, they are always dependent on color and shape characteristics, regardless of the detection and classification of traffic signs. It is a known approach to recognize color-based traffic signs [5],[6],[7] by identifying areas of interest for an image using a plain threshold or more progressive methods of image segmentation. The resulting regions are immediately marked as traffic signs or shift the suppositions of traffic signs (i.e. attention areas) into later phases. Since algorithms based on the RGB color space (Red, Green, Blue) are usually limited to varying light conditions in terms of adaptability, many researchers have transformed images into other color spaces, such as HSV (Hue, Saturation, Value) [8]. Yang et al. [9] introduced a color probability model based on Ohta color space to calculate probability maps for each color associated with traffic signs.

3.1.1.2 Segmentation by Colour

The colour of road traffic signs should be efficiently separated from environmental colours. Normally, image subdivision is a definitive process that determines the labels for each pixel in such a way that the equivalent labels have similar visual segmentation. The simplest segmentation method of an image can be applied by a specific threshold value for an image where each pixel that exceeds a certain threshold value has an appropriate label. Many studies have tried to identify the best colour alignment threshold. However, external influences could have a major impact on the achievement of color-based object segmentation techniques, such as light variations, shadows, and poor weather conditions. Recently, the colour threshold value is generally found as a pretreatment step to remove interesting areas [10]. Many studies attempted to reveal the effects of daily changes in light [11], in their experiment, they demonstrated an interesting technique in which they observed the red STOP color within 24 hours. The observation of the red colour component is more visible from 6:30 am until 21:00 pm. In this case the differences between the red, green and blue components δ_{RG} and δ_{RB} are the maximum, the red value is higher than 85% than the green and blue components. As stated in [11], color segmentation methods were proposed to correctly distribute red, green, and blue characters.

Another study focused on the development of a technique for detecting and identifying small sub-parts of small components of traffic signs [12]. In the first step, the algorithm used colour segmentation to locate red border regions. The segmentation model can be adjusted to different sensitivity conditions depending on the intensity level in order to avoid light sensitivity. Mean intensity values are rarely included when examining the upper edge of the input, which is generally corresponding to the boundary. These characters can be speculated for the bad weather conditions and choose the proper value. Gao et al. [13] employ the CIECAM97 color model, the images are first converted from RGB to CIE XYZ to LCH mode (Luminosity, Chrome, Tone). The study points out that the brightness values are quite similar for red, blue signs and background. Therefore, only the measurement of hue and chroma are considered for segmentation.

3.1.1.3 Shape-based Detection

Several approaches have proposed shape features for traffic sign recognition. Traffic sign shapes are generally circular, triangular, rectangular or polygonal. Hough Transform technique is probably the most common method of retrieving arbitrary patterns from an image. Wang et al. [14] developed a new ellipse detection technique for the detection of circular traffic signs, which are deformed by external force or shooting angle. Liang et al. [15] provided a set of templates corresponding to the shape for each traffic sign object category. Another interesting idea for determining interests is to practice a corner identifier and then a hypothesis regarding the position of common polygons by studying the relationship between angles. Paulo et al. [16] encountered both rectangular and triangular signs by initially engaging the Harris angle detector in an interesting region and then detecting the presence of edges in six predetermined dominance regions. The shape is defined by the composition of the adjustment ranges with angles.

3.1.2 Learning-based Detection Techniques

3.1.2.1 Detection based Machine Learning

Several investigations in the field of machine learning are applied to the problem of traffic sign recognition. They concentrate on the extracting of hand-crafted features for both the detection of a traffic sign and the classifying into one of several classes. The extracted features were employed to train classifiers such as support vector machines (SVMs), Bayesian classifiers, or logistic regression models. For instance, Baro et al [17] have used the AdBoosting [18] algorithm for traffic sign recognition. Wang et al. [19] used the Histogram of the Oriented Gradient (HOG) feature with the SVM classifier model to identify traffic signs and achieved excellent results. An enhanced common finder AdaBoost (CF. AdaBoost) algorithm was introduced by Liu et al. [20]. Chen et al. [21] suggested an accurate and efficient approach for the detection of traffic signs by combining AdaBoost with and support vector regression for discriminative detector learning. Although hand-crafted features have obtained a higher precision for traffic signs, it is observed that traditional recognition techniques have a much higher significance, however, they do not have robustness for the overall system.

3.1.2.2 Detection based Deep Learning

Nowadays, deep learning is widely employed in computer vision researches. Generally, there are two approaches to object recognition related to deep learning. On the one hand, it is based on a regional proposal and is also referred to as a two-stage approach (R-CNN) which significantly enhance the average detection accuracy in the challenge of visual object classes (VOC) [22]. R-CNN provides a remarkable improvement over previously published sliding window-based techniques, R-CNN uses selective search which is an unattended algorithm for generating feature extraction by CNN's separately. In the final step, R-CNN employs the SVM classifier to estimate the classes of objects. For better performance, it also applies linear regression to fine-tune the positions and sizes of detection boxes. After the spectacular effect of R-CNN, many new ideas have been introduced on CNN, such as Fast R-CNN [23], Faster R-CNN [24] and the spatial pyramid pooling network (SPP-Net) [25]. The precision and rapidity in object recognition of the above techniques enhanced dramatically and the fastest fame rate could be 15 fps. The other technique is depending on the regression method, which is an end-to-end learning model without classifiers and is also referred to as a one stage approach. YOLO algorithm (You Only Look Once), which is one of the faster object recognition algorithms [4]. It combines object detection and objects classification in a single convolution network. Although YOLO is no longer considered the most precise object detection method, it is a highly valuable solution when real-time detection is requested without sacrificing accuracy. YOLOv2 [4] and YOLOv3 [26] improve the original YOLO in several aspects. They include multiple network architectures, multiple connection box controllers, and several scales during training and detection processes. The SSD (Single Shot multibox Detector) [27] is another technique similar to the YOLO approach, which employs standard boxes and multi-scale feature mapping layers to increase accuracy. The majority of traffic sign detection techniques are image-based. An additional aspect of the study could be the analysis of traffic on/offline videos for traffic sign recognition using semantic representations [28] or based on a semi-supervised feature selection [29]. Regarding the deep learning approaches, the two-step strategies have advantages in terms of recognition accuracy and localization accuracy. However, the calculation efficiency is poor, and the process demands large amounts of time and resources. The single-stage methods are much accelerated by the unified network structures, although the process accuracy decreases. In addition, the amount of dataset is a major factor for deep learning-based methods.

In conclusion, traditional detection techniques have advantages in terms of accuracy. The algorithms based on deep learning can accomplish the robustness of an algorithm by self-learning to cope with different challenges. In this study, we present an end-to-end approach to recognize German traffic signs with a distance estimation inspired by YOLOv2. This can be performed faster and can be easily applied to a real-time system. In order to achieve precision, we consider the properties of traffic signs and CNN to enhance the network structure.

4 German Traffic Sign Recognition Dataset

The recognition of traffic signs is a real challenge of high automotive industrial interest. However, although some commercial approaches have been launched together with several studies on this issue, systematic unbiased comparisons of different approaches are still missing and benchmark datasets are not always freely available. The dataset used in this study is the German Traffic Sign Recognition Benchmark (GTSRB) [30]. The GTSRB is perfect for our goal as it is a large, organised and open source solution. GTSRB contains around 50,000 images in a total of German traffic signs, divided into 43 classes with major variations in terms of colour, shape, rotation, occlusion and weather conditions. A representative figure for each class is illustrated in Figure 2.

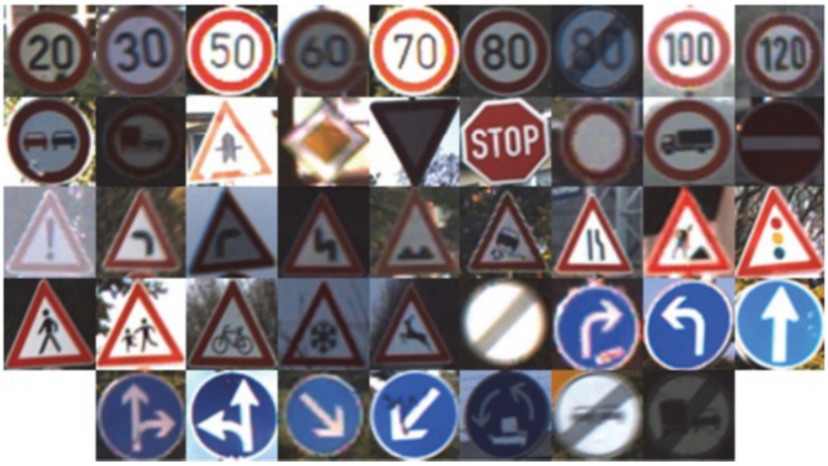


Figure 2 – A representative example of the 43 traffic sign classes of the GTSRB.

The dataset poses many challenges for classification, including varying light and weather situations, angle of view variations, motion blur, partial occlusions, physical defects, and other real variabilities (some samples considered hard to be classified). In this dataset, each image contains a traffic sign and a 10% border around the traffic sign itself. The traffic sign images are stored in PPM format. In addition, the resolution is not homogeneous, and the sizes vary between 15x15 and 250x250 pixels. The images are not generally square, and the traffic signs are not always centralized in the image. Figure 3 displays the distribution of test dataset classes from the GTSRB. As shown in Figure 3, the dataset consists of 43 classes with asymmetrical frequency.

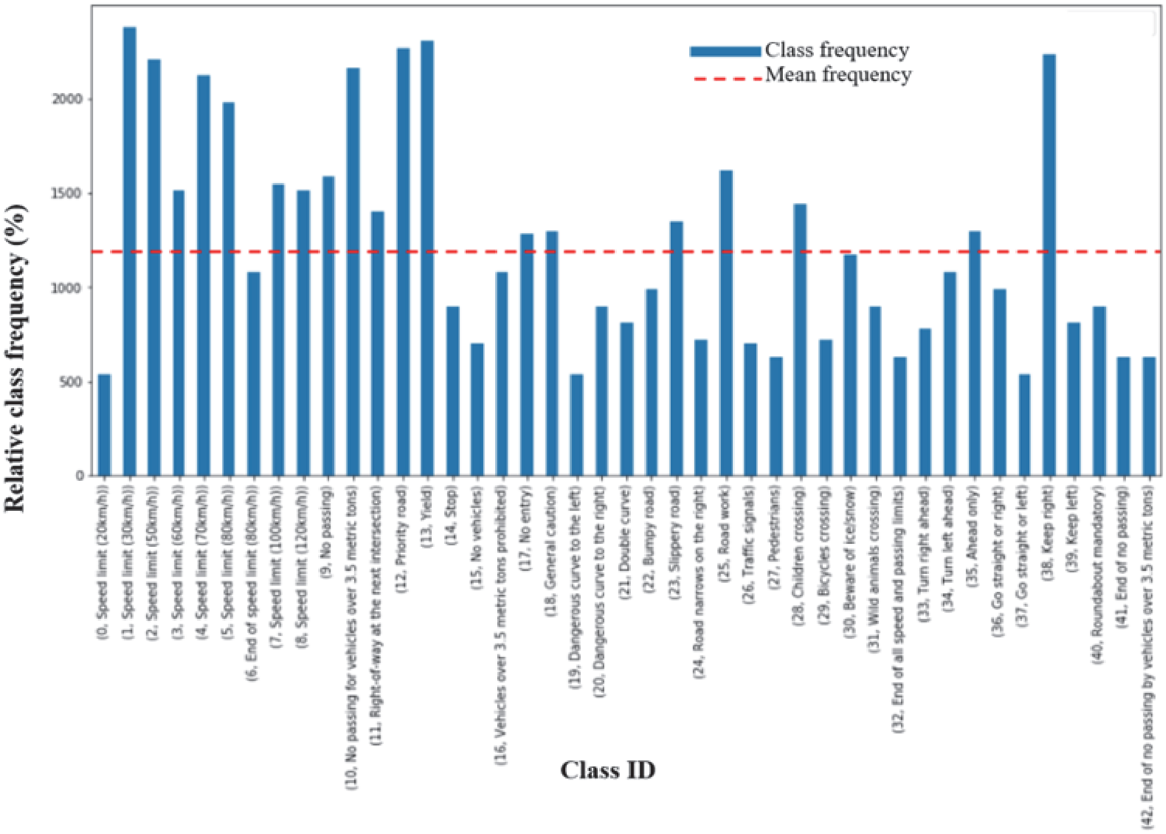


Figure 3 – Relative Class frequencies of the GTSRB dataset.

In order to balance the dataset, a part of the dataset was utilized. Overrepresented classes have been removed thus the frequencies of the various classes are roughly the same. However, it should be pointed out that the balancing of dataset might be problematic if the population dataset is unbalanced in themselves. If this is the situation, balancing the dataset can lead to inaccurate results.

5 YOLO Network Architecture

YOLO algorithm can view an image and then draw borders over what it perceives as identified classes. The question arises what makes YOLO different from other detection algorithms. Previous detection systems would use classifiers or localizers to perform the detection. The model would be applied to an entire image in various locations and scales. High score areas of the image are considered recognition. YOLO follows a completely different approach. A single neural network is applied to the entire image, it means "just look once". This network splits the image into different regions and provides a boundary boxes. These boundary boxes are weighted with the predicted probabilities. YOLO has several advantages over classifier-based systems. It looks at the entire image at test time, therefore its predictions are influenced by the global context in the image and not only in the particular region. This also provides prediction using a single neural network evaluation differently than other systems such as R-CNN, which employ thousands of bounding boxes to evaluate a single image, and then speeds it up to 1000x faster than R-CNN and 100x faster than Fast R-CNNN. In this study, we will use YOLOv2 for traffic signs detection and recognition. A single neural network is used by YOLO to predict bounding boxes and class probabilities directly from entire frames in an inference phase. It slices the input frame into S × S grids. For each grid cell, k boundary boxes and confidence values of boundary boxes, as well as C conditional class probabilities are estimated. In addition, each boundary box is expressed by the following tuple (x, y, Width, Hight, Cfd). Coordinates (x, y) represent the centre offset of the boundary box in comparison to the boundaries of the grid cell. Where Width and Hight correspond to the predicted width and height in terms of the entire image. Cfd represents the confidence value which is expressed as $P_r(Traffic_{sign}) * IOU_{Pred}^{G\_truth}$. When grid cell includes a portion of a ground truth box the value of $P_r(Traffic_{sign})$ will be 1, otherwise it will be 0. IOU indicates the intersection value over the union between the predicted boundary box and the ground truth box. These predictions allow us to derive the class-specific confidence value of individual bounding boxes and finally choose the high confidence scores of the bounding boxes in each grid cell in order to provide global predictions of a traffic sign in the input image.

In this study, a deep convolution network based on the end-to-end detection algorithm is developed to estimate real-time traffic sign detection. Figure 4 illustrates YOLOv2 architecture for the German traffic sign detection algorithm.

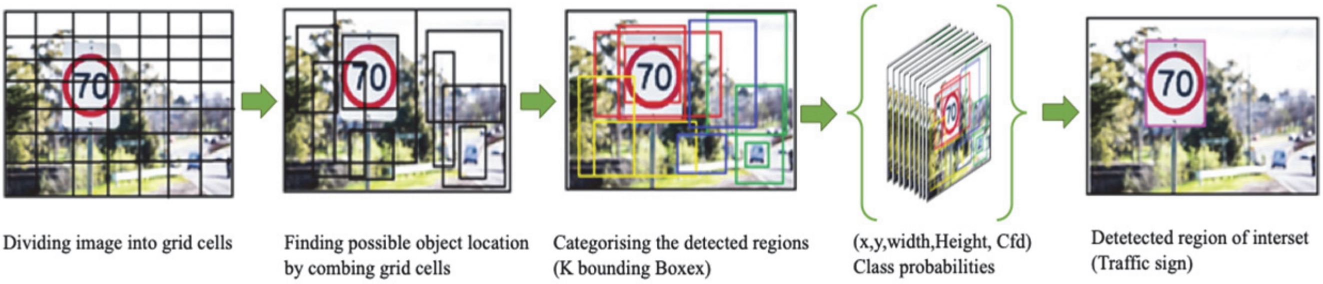

Dividing image into grid cells
Finding possible object location by combing grid cells
Categorising the detected regions (K bounding Boxex)
(x,y,width,Height, Cfd) Class probabilities
Detetected region of interset (Traffic sign)

Figure 4 – shows traffic sign detection algorithm based on YOLOv2 architecture.

For each grid, k boundary boxes are created corresponding to the previous anchor boxes. YOLO will add the λ_{Coord} parameter to enhance the loss for bounding box coordinate predictions when the bounding box covers the object (traffic sign). Optimum boundary box is determined by the value of IOU. In addition, YOLO includes the parameter λ_{Noobj} to reduce the loss from confidence predictions, as the box does not include the object (traffic sign).

5.1 Training Configuration

YOLOv2 has proposed a lightweight neural architecture that utilizes a convolution neural network for localization and classification. Its classifier part is referred as Darknet-19. This model mostly employs 3×3 filters, it has also used batch normalization, 19 convolution layers, and 5 max-pooling layers. It is initially trained on the ImageNet dataset (1000 classes) for 160 epochs at resolution 224x224. Default techniques are applied for data augmentation like saturation / colour etc. The Network is refined for 10 epochs onto resolution 448x448. In this study, the modified version of Darknet-19 was employed. In the modified model, the final convolution layer is removed, and replaced by three 3×3 convolution layers with 1024 filters each, followed with a final 1×1 convolution layer with the number of outputs required for detection are appended. In technical terms, 1×1 convolution kernel does not differ from any 3×3 kernel in the manner in which they are employed. However, they are conceptually differentiated. Usually, 3×3 convolutional are considered edge/feature detectors, while 1×1 kernel is able to combine only the activations of each feature vector. Interpretation of this kind of operation is that it simulates the impact of a fully connected layer that is applied to each feature vector. We have used 160 epochs, and during training the learning rate starts at 0.001 divided by 10 at 60 and 90 epochs and batch size of 32. The reason for the start with a lower learning rate is to avoid the model from diverging due to unstable gradients. Table 1 illustrates a complete description of the YOLOv2 model used in this work.

Table 1 – shows a YOLOv2 description configuration

Type	Filters	Size/Stride	Output
Convolutional	32	3 x 3	224 x 224
Maxpool		2 x 2 / 2	112 x 112
Convolutional	64	3 x 3	112 x 112
Maxpool		2 x 2 / 2	56 x 56
Convolutional	128	3 x 3	56 x 56
Convolutional	64	1 x 1	56 x 56
Convolutional	128	3 x 3	56 x 56
Maxpool		2 x 2 / 2	28 x 28
Convolutional	256	3 x 3	28 x 28
Convolutional	128	1 x 1	28 x 28
Convolutional	256	3 x 3	28 x 28
Maxpool		2 x 2 / 2	14 x 14
Convolutional	512	3 x 3	14 x 14
Convolutional	256	1 x 1	14 x 14
Convolutional	512	3 x 3	14 x 14
Convolutional	256	1 x 1	14 x 14
Convolutional	512	3 x 3	14 x 14
Maxpool		2 x 2 / 2	7 x 7
Convolutional	1024	3 x 3	7 x 7
Convolutional	512	1 x 1	7 x 7
Convolutional	1024	3 x 3	7 x 7
Convolutional	512	1 x 1	7 x 7
Convolutional	1024	3 x 3	7 x 7
Convolutional	1000	1 x 1	7 x 7
Avg Pool		Global	1000
Softmax			

6 Active Learning Approach

A supervised learning training set could be designed and prepared by a human expert annotating samples of data. However, it is not possible to employ this approach when the dataset becomes too large due to the time required and budget to perform the annotation. Therefore, it is necessary to select the most informative examples to annotate them. Active learning is a special form of machine learning where a learning mechanism is able to interact with the user to provide the expected outcome at the new unlabelled dataset. The principal assumption in active learning is that if a learning method can identify the specific dataset from which it will learn, it will be able to achieve better outcomes than traditional learning methods with much less dataset for training. The process of sub-setting the dataset is involving with an active learner which is supposed to learn from a particular strategy, which training subsets are suitable to maximize the accuracy of the model. Generally, there will be 4 different strategies for building these subsets of the dataset from the original training dataset:

- Random Sampling: the dataset is randomly sampled [31].
- Uncertainty Sampling: the examples where we are most uncertain about the class are chosen [32].
- Entropy Sampling: the examples whose class probability has the greatest entropy are considered [33].
- Margin sampling: the examples for which the difference between the most likely class and the second highest class are the smallest [37].

The integration of diversity into active learning in batch mode has been addressed by many researchers [34][35][36]. In this study, deep learning model was developed based on an active learning approach for traffic sign recognition and to examine its impact on classification performance. Whether it is possible to reduce the false positive rate by selecting the right training samples? The idea is to utilize an iterative algorithm for training using a small dataset of high priority, and then an active learning approach is used to select more data by labelling and model building, then the training a couple of roundup iteration then we manually override this similar data.

6.1 Active Learning Framework for Traffic Signs Recognition

In the first run of the model training, the framework has randomly selected ∼15% of the images (30,000) subset from the GTSRB dataset, then the chosen images will be annotated using the BBox-Label-Tool. This dataset is divided into training sets (70%) and test sets (30%). The model was trained on the training set and evaluated on the test dataset. In this case, when the accuracy is below the per-defined threshold, then we selected ∼ 0.4% of the images using the trained deep learning model. The images are not randomly selected. However, they are chosen based on the model performance on the unseen dataset. As shown in the figure 5.

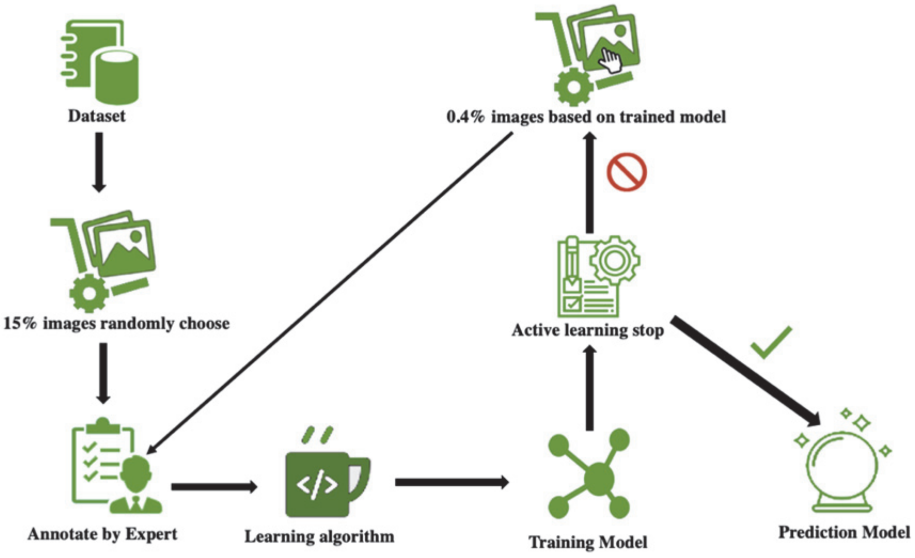


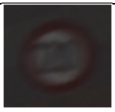
Figure 5 – shows suggested framework for an active learning framework for efficient development of deep model.

These images were not only chosen to represent a large variety of traffic signs. However, they were also selected to reflect several complex requirements. Therefore, the dataset will be split it into 3 different subsets:

- Positive Images-Set: it represents the high confidence images where the confidence value on traffic signs is higher than the 75% of defined threshold (we choose 45% in each iteration).
- Unsure Image-Set: it indicates the low confidence images where the confidence value on traffic signs is below the 75% of defined threshold (we choose 35% in each iteration).
- Negative Images-Set: it illustrates the images without prediction for which no traffic sign is expected (we choose 20% in each iteration).

In the iterative phase of the active learning proposed framework, the criteria for identifying the best candidates for the construction of a good model have been defined. These candidate images were carefully chosen not only to represent a large number of traffic signs, as well as to cover more challenging corner scenarios where the traffic signs are either partially occluded, faded traffic signs, distorted, or located far away in the background. For each training round, it chose ~ 0.4% of the original image set. These selections contained 45% of the images from the positive image-set, 35% from the unsure image-set and 20% from the negative image-set. Table 2 illustrates the selection criteria for images that are considered to be good candidates for establishing an effective model.

Table 2 – Shows examples of dataset samples

Class /Traffic Sign	Stop sign	20 km/h sign	70 km/h sign	120 km/h sign
Positive Images-Set (45%)				
Unsure Image-Set (35%)				
Negative Images-Set (20%)				

7 Distance Estimation

The estimated distance between a monocular camera position and the traffic sign is referred to as a regression problem [37]. Several strategies for estimating the distance between the target and the monocular camera position can be employed, such as the temporal method, which computes the distance based on the temporal sequence of object copies (e.g. visual odometry). The per-frame approach is another way in which the distance is predicted in the actual frame independently of previous frames (e.g. triangulation methods or depth estimation based on CNN techniques). In this study, taking into account that the size of the object is invariant and perpendicular distance calculation is required. Therefore, the area size of the object-based approach is used with a monocular camera. In this approach, the traffic sign is detected based on a deep learning model which predicts a boundary box indicating the position of the traffic sign element in the image frame. The size of these boundary rectangles is estimated at different locations and stored in a training dataset. By using curve fitting and optimization techniques, this dataset shows the nonlinear relationship between the area of the bounding boxes and the distance of the traffic sign. Therefore, the area, width and height of these bounded boxes are extracted for various ranges in order to estimate the nonlinear relationship between the area size of traffic sign and the distance (between camera and recognized sign). As shown in Figure 6, the distance shown is non-linear as a function of the area size. The SSD (Sum of Squared Difference) curve fitting technique is used to establish a relationship between the area size of the bounding box that contains the detected sign and the distance between that sign and the camera position and based on this relationship, it is possible to estimate the distance.

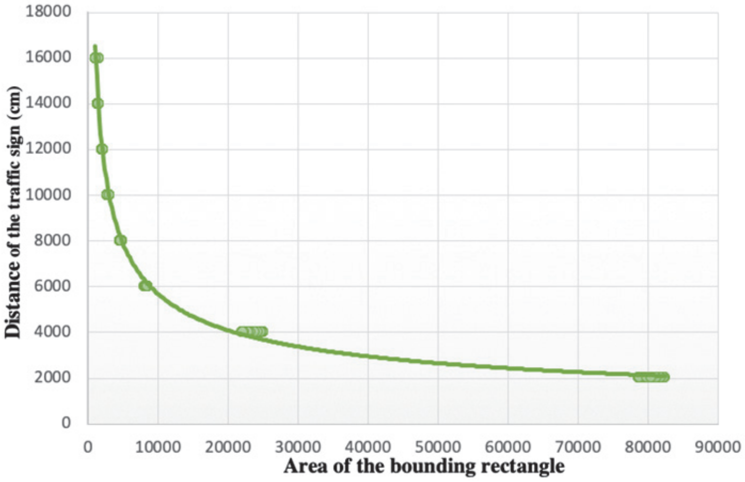
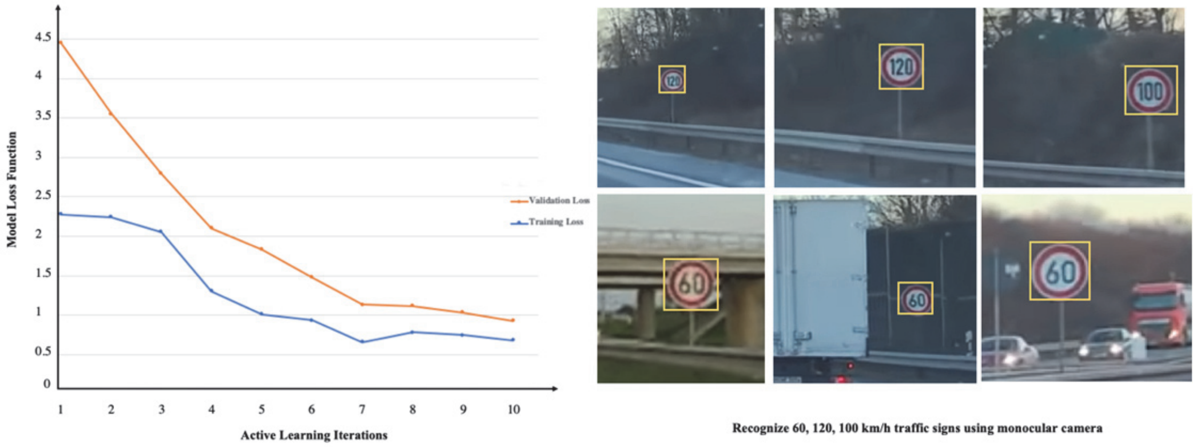


Figure 6 – shows Estimated distance plotted as a function of the area size of the bounding box

8 System Testing and Evaluation

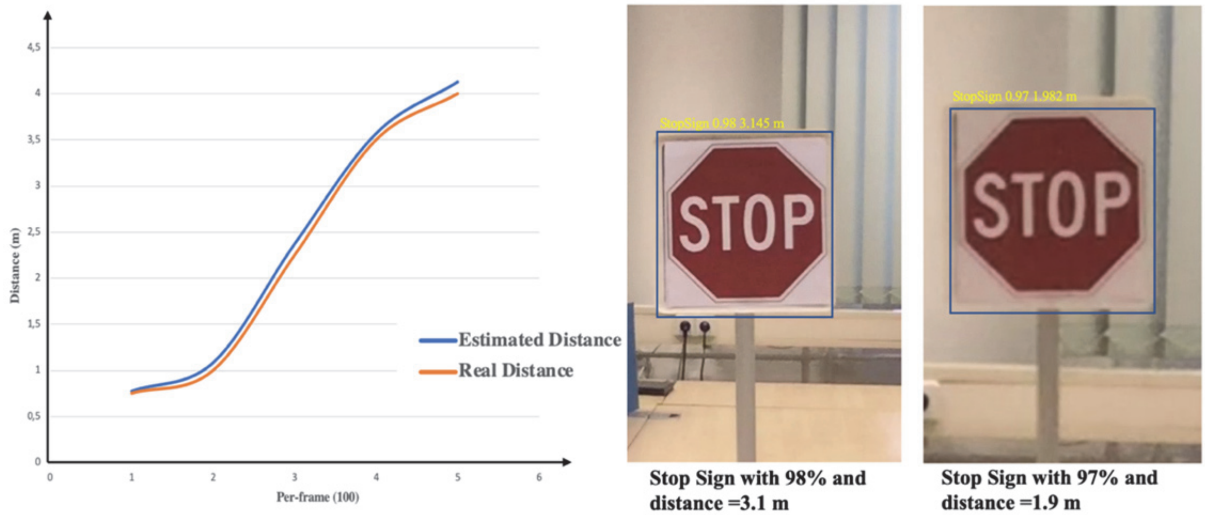
Ten iterations have been executed using our proposed active learning framework. At each iteration of the active learning approach, the actual training dataset consisted of the selected samples from all previous iterations and a new collection of samples specified by the suggested active learning framework (from 3 subsets of the dataset based on the confidence level). After each iteration, the performance of the trained model is evaluated, as it is shown in figure 7, the performance of the trained model is increased, and the value of model loss function is decreased.



(a) (b)

Figure 7 – (a) shows the performance of active learning model after 10 iteration, (b) shows detecting 60,120,100 km/h traffic sign on the test dataset.

The recognized traffic sign is then compared with the actual vehicle velocity in order to notify the driver to reduce or accelerate his vehicle velocity. In order to estimate a distance between traffic signs and monocular camera position, first, we need to detect the object (traffic sign) that we are interested in estimating its distance. Therefore, after successfully detecting the traffic sign, we can estimate the contour that corresponds to our object (traffic sign), which returns the boundary box containing the (x, y) coordinate width and height of the box (in pixels). Finally, we use rectangle sizes of the bounding box that are provided by the classifier to estimate the traffic sign distance. Figure 8 illustrates the error rate between the estimated distance and the actual distance which calculated using digital laser distance meter tool, and recognized stop sign with estimated distance in the lab.



Stop Sign with 98% and distance =3.1 m

Stop Sign with 97% and distance =1.9 m

Figure 8 – shows the error rate between the estimated distance (blue line) actual distance (red line)

In this study, the trained model relies on the YOLOv2 architecture including 19 convolution layers and 3 layers connected at the top of the convolution layers. This model provides approximately 97% mAP in the test dataset at high speed of approximately 60 ms/frame, which meets our real-time requirements. Comparing and evaluating our approach based on active learning techniques with traditional supervised learning algorithms on the labelled dataset. A new deep model with same YOLOv2 architecture has been trained on the GTSRB dataset using supervised learning algorithms (select samples from dataset randomly) for more than 500,000 iterations, this model produces about 92% mAP.   Figure 9 presents the model accuracy comparison between the active learning approach based on well-chosen samples and the supervised learning approach. As shown in this figure, our model can provide about 97% accuracy based on an active learning approach, compared to 92% who used all datasets to label and build the model.

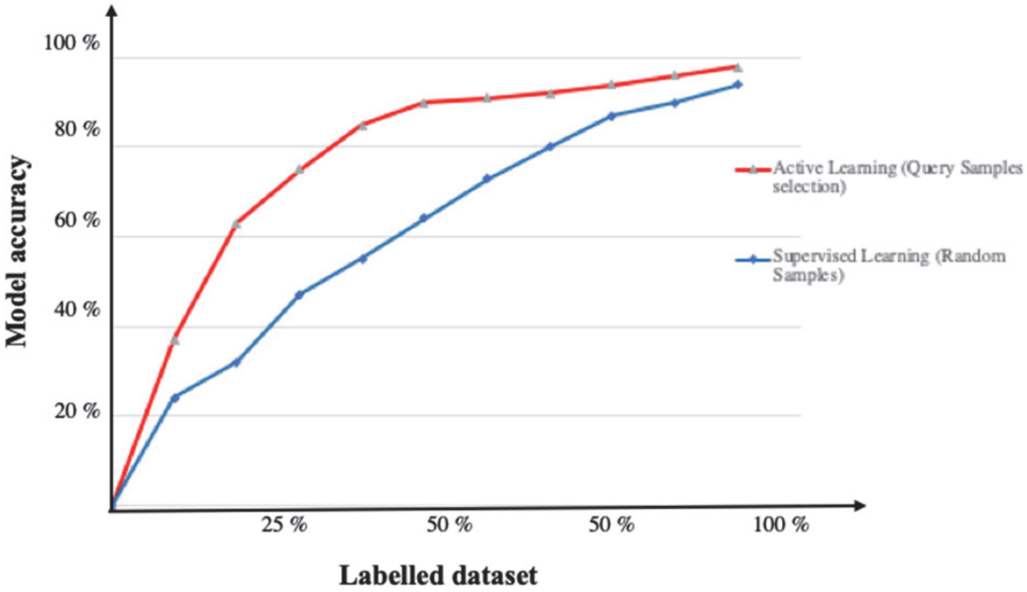


Figure 9 – shows 2 models learning curves illustrating that active learning approach demonstrates higher model accuracy with the same amount of label dataset compared to the supervised learning approach.

9 Conclusion and Future Work

In this study, deep learning and active learning techniques were deployed to provide smart identification and retrieval of the underlying information to be fed into the model at the time training phase. The model based on YOLOv2 is applied to recognize traffic signs with an average speed of approximately 60 ms/frame and with an accuracy of 97%. After successful recognition of the traffic sign, the distance between the traffic sign and a monocular camera is estimated based on the ratio between the area size of the boundary box containing the traffic sign and its actual size at different distances. Furthermore, our model, based on the active learning approach, achieved higher model accuracy with the a few amount of labelled dataset when compared to the supervised learning approach. YOLOv2 is a single neural network which directly provides bounding boxes and probabilities from a whole image in only one evaluation. It provides high precision recognition of traffic signs and achieves a higher recall rate with fewer false positives.

In some situations, the detected boxes were not exactly located on the traffic signs overall and were often not fixed around the traffic signs. This problem should be addressed in future improvement work as it is important to have precise bounding boxes for accurate recognition of traffic signs and distance estimation. In the future work, the calculated distance should be improved, with minimized errors. Therefore, the depth information should be estimated in the meantime with objects recognition using a single monocular camera and a light neural architecture to predict pixel-wise depth map. Furthermore, it will be necessary to improve the current framework and tools that allow for such a development and to demonstrate how they can be applied to tackle a wide range of real-world challenges.

References

[1] S. B. Wali *et al.*, “Vision-Based Traffic Sign Detection and Recognition Systems: Current Trends and Challenges.,” Sensors (Basel)., vol. 19, no. 9, May 2019.

[2] J. Li and Z. Wang, “Real-Time Traffic Sign Recognition Based on Efficient CNNs in the Wild,” IEEE Trans. Intell. Transp. Syst., vol. 20, no. 3, pp. 975–984, March. 2019.

[3] A. Shustanov and P. Yakimov, “CNN Design for Real-Time Traffic Sign Recognition,” Procedia Eng., vol. 201, pp. 718–725, 2017.

[4] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.

[5] L. D. Lopez and O. Fuentes, “Color-Based Road Sign Detection and Tracking,” in Image Analysis and Recognition, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 1138–1147.

[6] C. Bahlmann, Y. Zhu, Visvanathan Ramesh, M. Pellkofer, and T. Koehler, “A system for traffic sign detection, tracking, and recognition using color, shape, and motion information,” in IEEE Proceedings. Intelligent Vehicles Symposium, 2005., 2005, pp. 255–260.

[7] S. B. Wali et al., “Vision-Based Traffic Sign Detection and Recognition Systems: Current Trends and Challenges,” Sensors, vol. 19, no. 9, p. 2093, May 2019.

[8] Wang, G.Y.; Ren, G.H.; Quan, T.F. A traffic sign detection method with high accuracy and efficiency. In Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE), Hangzhou, China, 22–23 March 2013; pp. 1426–1429.

[9] Yang, Y.; Wu, F.C. Real-time traffic sign detection via color probability model and integral channel features. In Proceedings of the 6th Chinese Conference on Pattern Recognition (CCPR), Changsha, China, 17–19 November 2014; pp. 545–554.

[10] A. Ruta, Y. Li, and X. Liu, “Real-time traffic sign recognition from video by class-specific discriminative features,” vol. 43, no. 1, pp. 416–430, 2010.

[11] M. Benallal and J. Meunier, “Real-time color segmentation of road signs,” Electrical and Computer Engineering, 2003. IEEE CCECE 2003. Canadian Conference on, vol. 3, pp. 1823–1826 vol.3, May 2003.

[12] L. Estevez and N. Kehtarnavaz, “A real-time histographic approach to road sign recognition,” Image Analysis and Interpretation, 1996., Proceedings of the IEEE Southwest Symposium on, pp. 95–100, Apr 1996.

[13] X. Gao, L. Podladchikova, D. Shaposhnikov, K. Hong, and N. Shevtsova, “Recognition of traffic signs based on their colour and shape features extracted using human vision models,” Journal ofVisual Communication and Image Representation, vol. 17, no. 4, pp. 675–685, 2006.

[14] Wang, G.Y.; Ren, G.H.; Wu, Z.L. A fast and robust ellipse-detection method based on sorted merging. Sci. World J. 2014.

[15] Liang, M.; Yuan, M.; Hu, X. Traffic sign detection by ROI extraction and histogram features-based recognition. In Proceedings of the International Joint Conference on Neural Networks, Dallas, TX, USA, 4–9 August 2013.

[16] C. Paulo and P. Correia, “Automatic detection and classification of traffic signs,” in Image Analysis for Multimedia Interactive Services, 2007. WIAMIS ’07. Eighth International Workshop on, June 2007.

[17] X. Baro, S. Escalera, J. Vitria, O. Pujol, and P. Radeva, “Traffic sign recognition using evolutionary adaboost detection and forest-ecoc classification,” IEEE Transactions on Intelligent Transportation Systems, vol. 10, pp. 113–126, March 2009.

[18] P. Viola and M. Jones, “Fast and robust classification using asymmetric adaboost and a detector cascade,” vol. 14, pp. 1311–1318, 2002.

[19] Wang, G.Y.; Ren, G.H.; Wu, Z.L.; Zhao, Y.Q.; Jiang, L.H. A robust, coarse-to-fine traffic sign detection method. In Proceedings of the 2013 International Joint Conference on Neural Networks, Dallas, TX, USA, 4–9 August 2013; pp. 754–758.

[20] C. Liu, F. Chang, and Z. Chen, “Rapid multiclass traffic sign detection in high- resolution images,” IEEE Transactions on Intelligent Transportation Systems, vol. 15, pp. 2394–2403, Dec 2014.

[21] T. Chen and S. Lu, “Accurate and efficient traffic sign detection using discriminative adaboost and support vector regression,” IEEE Transactions on Vehicular Technology, vol. 65, pp. 4006–4015, June 2016.

[22] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

[23] Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.

[24] Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 1137–1149.

[25] He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2015, 37, 1904–1916.

[26] Redmon, J.; Farhadi, “YOLOv3: An Incremental Improvement,” Apr. 2018.

[27] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C. SSD: Single shot multibox detector. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.

[28] Chang, X.J.; Ma, Z.G.; Yang, Y.; Zeng, Z.Q. Bi-level semantic representation analysis for multimedia event detection. IEEE Trans. Cybern. 2017, 47, 1180–1197.

[29] Chang, X.J.; Yang, Y. Semi-supervised feature analysis by mining correlations among multiple tasks. IEEE Trans. Neural Netw. Learn. Syst. 2017, 28, 2294–2305.

[30] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, “Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition”, Neural Networks, vol. 32, pp. 323– 332, 2012.

[31] H. Woo and C. H. Park, “An Efficient Active Learning Method Based on Random Sampling and Backward Deletion,” Springer, Berlin, Heidelberg, 2013, pp. 683–691.

[32] Y. Yang and M. Loog, “Active learning using uncertainty information,” in 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 2646–2651.

[33] A. Holub, P. Perona, and M.C. Burl. Entropy-based active learning for object recognition. In Computer Vision and Pattern Recognition Workshops, 2008. CVPRW’08. IEEE Computer Society Conference on, pages 1–8. IEEE, 2008.

[34] J. Zhou and S. Sun, “Improved Margin Sampling for Active Learning,” Springer, Berlin, Heidelberg, 2014, pp. 120–129.

[35] J. Zhang et al., “A Real-Time Chinese Traffic Sign Detection Algorithm Based on Modified YOLOv2,” Algorithms, vol. 10, no. 4, p. 127, Nov. 2017.

[36] H. Irshad, Q. Mirsharif, and J. Prendki, “Crowd Sourcing based Active Learning Approach for Parking Sign Recognition,” Dec. 2018.

[37] Quoc V Le, Jiquan Ngiam, Adam Coates, Abhik Lahiri, Bobby Prochnow, and Andrew Y Ng. On optimization methods for deep learning. In Proceedings of the 28th International Conference on International Conference on Machine Learning, pages 265–272. Omnipress, 2011.