

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

VEHICLE DETECTION USING VISUAL ATTENTION BASED SALIENCY MAP AND CONVOLUTION NEURAL NETWORK

Jenish Modi^{*1}, Bhaumik Vaidy^{a2} & Chirag Paunwala³

^{*1}PG student, EC Department, SCET, Surat, India.

²Research Scholar, GTU, Ahmedabad, India.

³Professor, EC Department, SCET, Surat, India.

ABSTRACT

Accurate detection of the vehicle is important for self-governing vehicle systems. In this paper, we are discussing vehicle detection algorithm using visual attention based saliency map and deep learning. In addition, it uses the Convolutional Neural Network (CNN) to extract and detect images features. To improve the detection accuracy and reduce detection time, saliency map is used and it identifies the region-of-interest in the visual image. A saliency map containing the vehicle location is generated. Then CNN is used to classify the class of object. The proposed algorithm has been evaluated on variety of images of real environments. The experimental results show the reduce vehicle detection time of the proposed algorithm and reduce different challenges like rotation, different color, and scaling of the object.

Keywords: Convolutional Neural Network, saliency maps, visual attention, visual search, vehicle detection

I. INTRODUCTION

Object detection is to detect objects from a known class, like cars, faces or people from an image. Every object class has its own features that helps in classifying the object. Object detection can be used as an applications for face detection, vehicle detection, person counting, and security. In recent years, object detection is a challenging research area. It is difficult to develop object detection system which has high detection accuracy irrespective of the quality of input images. A traditional object detection is a combination of object localization and classification. Many old image classification algorithms follow this process: It starts with image sensing, image pre-processing, object segmentation, feature extraction and object classification. In this process, relevant features must be extracted manually from the image. In transport system, vehicle detection is important task of computer vision.

Many segmentation methods are available like: Thresholding method, edge-based segmentation method, region-based segmentation method, and clustering based segmentation method, Artificial Neural Network (ANN). These methods are used for locating the object in an image. Saliency is one method for image segmentation. A saliency map shows each pixel's exclusive quality. The goal of a saliency map is to change the representation of an scene into something that is easier to analyse [1]. Image segmentation is typically used to identify the outlines in images. The result of image segmentation in a region is similar with respect to some feature, such as color, intensity, or texture. The saliency maps are very robust to noise [3].

During object detection, many challenges like illumination variations, rotation, object shadowing, scaling and occlusion occur. The object is detected by extracting relevant features manually from the image [2]. But manually extracted features cannot transfer from one application to another application. Therefore, deep learning models can be used to extract features automatically which improves the accuracy of object detection.

Now a day “Deep Learning” has been one of the trending technique. It is a machine learning technique that learns features directly from data. Data can be text, sound or images. Deep Learning based algorithms fork the feature extraction step completely. Architectures of the neural network are useful for deep learning. Convolutional Neural Network (CNN) is a most popular method of deep learning. The term “deep” usually refers to the number of hidden layers in the neural network. Traditional neural networks only contain 2-3 hidden layers, while some recent deep networks have as many as 150 [2]. However, Deep learning has become very popular recently because it can be

extremely accurate without need to extract feature manually in an image. There is no need to understand which features are the best representation of the object. Deep learning models will require large amount of data and it will take long time to train.

The proposed method is to identify the vehicles location which adopts a saliency map that generates salient region present in an image. The deep learning framework, i.e. convolutional neural network, is used to classify the vehicle form an input image. The proposed method is good under varying condition like object rotation, different color and scaling.

II. RELATED WORK

Most object detection methods used different background subtraction methods or different object segmentation methods. Every method has its pros and cons. For example, the advantage of thresholding method is that it does not require previous information and it is simplest method but at the sometime it is highly dependent on peaks and spatial details. Similarly, edge based detection method provides better contrast between two objects but when there are too many edges in one image, at that time pixels can be wrongly classified in different objects. Tiantian Wang, et al. [1], proposed the object detection method based on image saliency which uses local features that highlights the object. Then, input image converted into HSV color space, and then hue component is used to identify the object using the color histogram method. Similarly, Jacob Gildenblat et al., describe saliency map generation method on the simple image using histogram back projection. Laurent Itti, et al. [3] proposed saliency-based visual attention model. Now a days, to improve the detection accuracy, researchers have combined many features like intensity, color, orientation to generate saliency region which is useful for the object location information.

After finding the location of object, it need to be classified. The advantage of Artificial Neural Network (ANN) method is, it efficiently handled high noisy inputs and computation rate but drawbacks of method is over-fitting and training is time consuming. Support Vector Machine (SVM) removes problem of over-fitting and reduction in computational complexity but transparency of the result is low and more time is consumed in training. Vijay John, et al., proposed classification using CNN algorithm and compared it with baseline algorithms. Now a day's, to improve the detection accuracy, researchers have gone to deep learning algorithm which is used to classify the object.

III. SALIENCY MAP GENERATION AND CLASSIFICATION METHOD

A. Visual attention based saliency map model

In this method input in the form of static color images. Using dyadic Gaussian pyramids [3] nine spatial scales are created, which progressively filtered and sub-samples the input image, producing horizontal and vertical image reduction factors ranging from 1: 1 (zero scale) to 1: 256 (scale eight) in eight octaves.

Each feature is calculated from a set of "center-surround" linear operations. This method is sensitive to discontinuities and good for detecting places of object that distinguish itself from the surrounding environment. Then finds center-surround difference between fine and coarse scales: here center scale is c

$\{2, 3, 4\}$, and the surround scale is $s = c + j$, with $j \in \{3, 4\}$. The across-scale difference between two maps, denoted “ Δ ”. The finer scale result is generated by interpolation and point-by-point subtraction. Feature extraction not only for c scale but also for $j = s - c$, by including ratios of different size between the center and surrounding regions.

In mammals, the first feature map is intensity contrast which means neurons sensitive either too bright centers on dark surrounds or too dark centers on bright surrounds. Here, both types of sensitivities are calculated in a set of six maps $I(c, s)$, with $c \in \{2, 3, 4\}$ and $s = c + d$, $d \in \{3, 4\}$:

$$\bar{a} = \frac{a - b}{a + b} \quad (1)$$

Similarly, the second feature map is color spaces. In this feature map center neurons excited by one color and surrounding neurons excited by another color.

□ à (2)

Third feature map is based on orientation. Local orientation information is obtained from I using oriented Gabor pyramids $O(c, \Lambda)$, here $[0..8]$ represents the scale and $\Lambda \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ is the preferred orientation. Orientation feature maps, $O(c, s, \Lambda)$, encode, as following:

$$\tilde{a} \tilde{\Lambda} \square \tilde{a} \tilde{\Lambda} \tilde{a} \tilde{\Lambda} \quad (3)$$

Above three feature maps, \tilde{a} , $\tilde{\Lambda}$, and $\tilde{a} \tilde{\Lambda}$ normalized individually that similar features compete well for saliency. The three conspicuity maps are normalized and summed into the final input \tilde{a} to the saliency map:

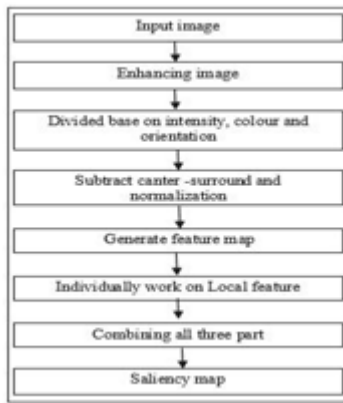


Figure 1. Flow chart of visual attention based saliency map model [3]

$$\square - \quad (4)$$

Classification using CNN

Result of saliency map become the input of CNN and it classifies a class of the object. The deep learning is a part of machine learning technique. For extracting feature and classification many layers are used. In the CNN framework the bias and weights are used in back-propagation algorithm to learn at each layer. Into the many layers, the lower layers are used to learn color, edge and texture feature, while the higher layers learn contour features. The weights and bias got in the last layers are used to classify the class of object. Different layers of CNN architecture are the convolutional, rectified linear units, pooling, fully connected (FC), drop out, and loss layers. The first layer is convolutional layers, which generate a feature map using convolution with the previous layers value and current weights. Then next is activation function. Here, relu layers are used as an activation function which increases the non-linearity. Then take translation invariance layers that is called as pooling layer, which reduces the size of the image. Finally, the fully connected layers are connected, which take the number of neurons as per requirement. Then solve the over-fitting problem using drop-out layers. Last but not the least, the loss layers are connected, which evaluate the difference between the network output and the target. This process is repeated many time.

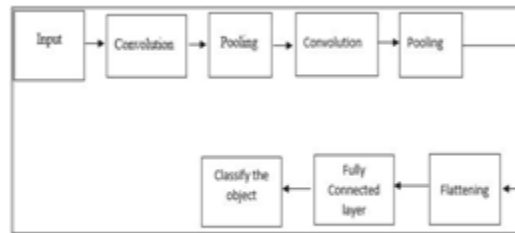


Figure 2. Block diagram of classification using CNN

The proposed algorithm consists of two convolutional layers, two pooling layers, one flattening layer and one FC layer. Details are given as follows:

$64 \times 64 \times 3$ is, the size of the input image. The first convolutional layer has 32 filters with 3×3 kernel size and relu layer. After that take maximum pooling layers, which is done with 2×2 filter size and 1 stride. The second convolutional layer has 64 filters with 3×3 filter size, and maximum pooling layer with filter size 2×2 and stride 1 and the third layer is flattening layer which converts matrix in to column vector. Then one fully connected layer, which contains 128 neurons with relu function. The one output neurons correspond to the two classes: auto rickshaw or no auto rickshaws and output.

Similarly, the input layer and the output layer of the multi-class CNN architecture in our proposed algorithm consists of two convolutional layers, two pooling layers, one flattening layer and one fully connected layer.

C. Propose method

Figure 3 shows flow of proposed method. In this method first

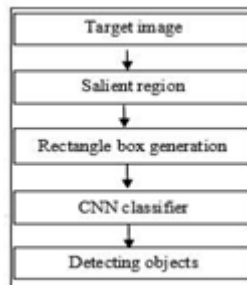


Figure 3. Flow chart of propose method

we take the color image as an input and then applied visual attenuation based saliency map [3] algorithm for generation salient regions. This salient regions represent the region of the interested object present in an image. By using this salient regions we draw the rectangle boxes on that objects. Then these regions are separated from original image and given to CNN to classify the class of object. Separation of region, reduces process time.

IV. IMPLEMENTATION AND RESULTS

A. Dataset:

Our algorithm was implemented with anaconda software and python language on Intel(R) Core(TM) i3-5005U CPU @ 2.00GHz, 2 Core(s), 4.00 GB RAM. Here, IIIT Hyderabad dataset [5] is used for in proposed method.

B. Comparative Analysis:

In this analysis we compare different method of saliency map generation as shown in table-1.

Table I. comparison of performance of different method of saliency map generation

Original image	Local Feature base method [1]	Back projection method	Visual Attention Model [3]
			
			
			
			

After performing the different method on saliency map generation, visual attention base method is accurate as compare to other two methods. Four database given as running automobile, parked cars, single auto-rickshaw and oriented rickshaw. Table-1 represents the result of three methods on different images. Table-1 shows that third method gives good result compared to other two. In first image, Method-1 is inaccurate for interested region and second method gives irrelevant information.

C. Performance Analysis

In this analysis we take different size of training and testing image set and compared accuracy and time of that image set using CNN classification algorithm.

Table-2 has different image size for data set and shows validation accuracy and training time. If we consider single image with different sizes as input then binary classifier is used for object detection and then obtained result shows 93.00 % accuracy and 467 sec training time after one epoch and 88.36 accuracy and 1450 sec training time after three epoch. If training and testing image size is reduced then the result would be after first epoch is 90.00 % and after third epoch is 95.00 % accurate and it take 116 and 121 seconds for training time respectively.

When we change the different parameter of CNN model like input image size, filter size, number of epoch, and number of neurons detection accuracy changes. Here all results are perform on binary class classification using same number of training and testing images. Here, we had taken 160 images for training and 20 images for testing. Therefore, each class has 80 images for training and 10 images for testing.

After performing variation in input size of image we can observe that increasing of size of input image leads to increase in time of training and testing periodically. Similarly, if we increase size of filter then training time, testing time and accuracy also increased. If number of epoch improves i.e.1, 2, ... likewise , then training and testing time gradually increasing.

TABLE II. EXPERIMENTAL RESULTS OF CNN CLASSIFICATION







Training input images	Testing Input images	Epoch	Prediction image	Validation accuracy (%)	Training Time (second)
600	100	1	Auto-rickshaw	93.00	467
		3		88.00	1450
160	20	1	Auto-rickshaw	90.00	116
		3		95.00	121
160	40	1	Bus	87.50	100
		3		90.00	294
800	200	1	Bus	67.46	348
		3		70.44	976

Furthermore, if we can increase number of neurons in fully connected layer then we training time and testing time both are increased but validation accuracy was reduced after certain level of increase in neurons. Here, we observed training time, testing time and accuracy because time and accuracy both are important parameter for measuring the performance of object detection. If system wants more accuracy then it takes high amount of time, hence it is not useful for real time application.

Table-3 show the performance of imposes algorithm is various challenges like rotation, scaling and change in color of the object.

Here, describes the result of the proposed system using saliency map and without the use of saliency map. The image is directly given to CNN classifier which classifies the object. The output represented in the figure. 4. While in figure 5 it describes the result of each step of proposed system using saliency map. In this case, each rectangle box was given to separately in CNN classifier.

Table III. different challenges occurred during object detection

Challenges	Original	Output
Rotation		
Colour change		
Scaling		

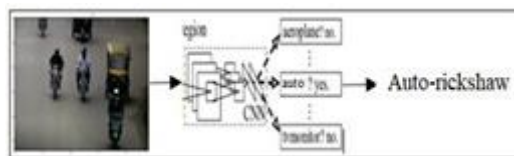


Figure 4. Result of proposed system without saliency map

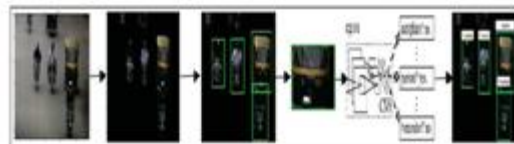






Figure 5. Result of proposed system using saliency map

Table-4 gives the time information during the experiment on the images. Consider image given in the first row and measure time with and without saliency map that is 20.95 and 3.53 second respectively. Similarly, the time taken is 3.43 and 3.18 second without and use of saliency map respectively.

Table IV. time of with and without saliency map of image

Input images	Saliency map result	Time(second)	
		Without saliency map	With saliency map
		20.9563	3.5341
		3.4319	3.1892

window technique. This will require more windows to be processed so it will take more time to detect the object while in propose method saliency maps so it will reduce number of windows to be processed. That will reduce time taken for detection of object on same image. The comparison of time taken by both methods is shown in table-4.

V. CONCLUSION

In this paper, vehicle localization and classification algorithm are proposed using saliency map generation and CNN respectively. Saliency maps are used to segment region of interest and the convolutional neural network are used to classify the class of object. Generate the saliency maps using the visual attention base method. These generated saliency maps output given to CNN classifier. We applied the proposed algorithm to IIIT Hyderabad dataset and the experiments proved a less vehicle detection time. Additionally, we performed varying parameter evaluation of CNN classification algorithm and concluded that the CNN -based vehicle detection is good using saliency map generation method.

REFERENCES

1. Kaur, Dilpreet, and Yadwinder Kaur. "Various image segmentation techniques: a review." *International Journal of Computer Science and Mobile Computing* 3, no. 5 (2014): 809-814.
2. V. John, K. Yoneda, Z. Liu and S. Mita, "Saliency Map Generation by the Convolutional Neural Network for Real-Time Traffic Light Detection Using Template Matching," in *IEEE Transactions on Computational Imaging*, vol. 1, no. 3, pp. 159-173, Sept. 2015.
3. Itti, Laurent, Christof Koch, and Ernst Niebur. "A model of saliency-based visual attention for rapid scene analysis." *IEEE Transactions on pattern analysis and machine intelligence* 20, no. 11 (1998): 1254-1259.
4. Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghahfouri, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I. Sánchez. "A survey on deep learning in medical image analysis." *arXiv preprint arXiv: 1702.05747* (2017) ELSEVIER.
5. Dataset link: http://cvit.iiit.ac.in/autorickshaw_detection/.