Proceedings of the International Conference on Communication and Electronics Systems (ICCES 2018)
IEEE Xplore Part Number:CFP18AWO-ART; ISBN:978-1-5386-4765-3

Developing an Algorithm for Scene Understanding and Decision Making Based on Visual Perception

Yashwantrao P Nimbalkar
Department of Electronics and Telecommunication
MIT Academy of Engineering,
Pune, India
yashwantrao.kknimbalkar@gmail.com

Shridhar A Khandekar
Department of Electronics and Telecommunication
MIT Academy of Engineering,
Pune, India
sakhandekar@etx.maepune.ac.in

***Abstract*— Trending is the era of artificial intelligence, which now a day plays very crucial roles for performing daily repetitive as well as decision making complex tasks. Sensing the dynamic scenario and making decision based on visual perception is blooming a bit. In this we propose an algorithm in which we divide the image in two categories frame of reference and the visual objects. Frame of reference gives a platform on which the object is placed that deals in making environment decision. We use faster RCNN algorithm for object and background detection. In co-relation of object, there motion and environmental condition, for making decisions Naives bayes algorithm is developed which is used to sense the surrounding and automatically make a decision based on visual perception and coordinate needs. At the end we analysis the experimental results such as training accuracy, total losses function which has pretty fine reliability and accuracy.**

Keywords—Scene understanding, faster region based CNN.

1. Introduction

Giving a vision to a machine is a challenge to human world. A lot of research is going in this field to provide a visual cortex to a machine. One of the problems of visual scene understanding entails recognizing the semantic constituents of a scene and the complex interaction that occurs between them and Inter-object blocking, representing complex object interaction as a physical residual and geometrical reference in an object. Using an algorithm for scene understanding which requires object detection and making co-relation between object and surroundings of an image is at the heart of this problem. The introduced models provide the state of art in understanding and investigating the object detection output for related image retrieval and classification tasks. This model allow you to integrate with visual references and multi-object tracking from a supervisor, which only uses the objects of multiple video as an input. For many other methods, a clear perspective makes sense, and thus tracking objects are enabled for a partially extended period, or objects whose full scales have never been observed. In addition to this, the joint scene tracklet model improves performance for more than one frame [1]. Semantic information in abstract image created from clip art collection which offers several advantages over the real images. It allows the study of high-level meaningful information directly in the study, because it eliminates the hand-labelling of low-level objects, attributes and relationship detectors, or real images [2]. Multiple CNN structures with various data-fusion strategies and wet-sharing schemes are proposed to learn both local and temporary connectivity from this motion channel [3].

A potential object and extracting it from a complicated scene is a problem, Solve the problem by using faster region based convolutional neural network which can locate a potential object in a scene with a few training samples [4]. Combining different collected feature maps improve the ability of region of interest (ROI) to draw more detailed features. Use the coordinates of ground object and many detailed information, such features information about free space, point cloud density use to build bounding box to find object pose [5]. However, the existing methods working for single object detection cannot directly apply for multiple object detection. Show the multiple object detection by exhibiting a framework that combine with region based convolution neural network and deformable part based model [6]. Introduce a region proposal network used to predict the probability of objectness score and object bound. The Region proposal network merge with faster R-CNN which share common single convolution features for object detection [3].

We wish to introduce a faster Region based Convolution neural network for locating a potential object in an image frame. We mainly focus on developing faster R-CNN model which is different from the existing method.

A. Object Classifiction

Convolution Neural Network required a large amount of data to train a classification model. Overcome this problem by using a transfer learning paradigm uses a pre-trained CNN with sufficient training data [7]. Presents a methodology based on score level fusion which uses kernel fisher analysis for extraction of feature phase and CNN model for developing feature maps [8]. These methods proved to be useful in classifying object with complex features. Develop a model using autowave metric and inary morphology with acute low false alarm rate. Debris particles are separated using autowave metric distance [9]. One of the drawbacks of convolution is that it fails in extracting correct geometry of the required object.

Solve this problem by introducing context dependent methodology which uses neighbourhood points to extract the object geometric features [10].

However, as the complexity of an image increases its representation also became complex. To deal with high level representation, proposed an object-to-class distance model for scene image [11]. It increases description of a particular space by using lower dimensional object bank. Present a different approach which correlate with different scenes to minimize the error in training samples [12].

B. Deep Learning

Prolong training time of deep learning network and limitation of it to multicore GPU which is affordable for high end servers is a major drawback. Develops a deep learning technique uses SoC implementation for lower cost gadgets [13]. It's having multi parallel algorithm which ease complex functionality of tradition deep learning network. Propose a SIFT vector feature model and DNN model for analysis of discriminating optimal feature for classifying objects [14]. Uses hierarchical deep multi-task learning (HD-MTL) algorithm which can combine two prudent regulation rules for effectively controlling inter-level error propositions and they can jointly provide an end-to-end approach to learning more intensive CNN features. This approach is having intensive learning algorithm which can effectively adapt to new training images using deeper CNN and tree classifier for new object classes [15].

Dense correspondence based transfer learning technique extract extensive features of a scene by using convolution neural network by creating concise and effective feedback through cross-domain metric learning and subspace alignment for cross-domain recovery and fetching the interpretation from best equivalent image to test image by using cross-domain dense correspondences and a probabilistic Markov random field [16]. Collection of grid to distinguish between active and still object and modelling information imposes relevant concessions on objection which is able to determine Dynamic Object, Parked Car, Urban Infrastructure and Able to Determine the Status of Cells occupied by Building [17]. Two personalized CNNs use objects on input data and classify their categories. RGB data can be found on non-invasive features, which are necessary in strict, disordered and altered environments [18]. Visual images are non-interactive, and offer Fuzzy Qualitative Rank Classifier (FQRC) to handle the above issues [19]. Proposed FQRC makes provision for ranking explanations rather than binary decision. Using qualitative and public view datasets, qualitative and quantitative duration assessments have shown the effectiveness of our proposed methodology in non-interactive exchange image modelling. A possible potential probabilistic Expectation-Maximization (EM) formulation, in which two mutually qualified steps are performed in a specific way, automatically provides supplemental information to each other [20]. A framework that integrates support vector machine-based trace detection with a trail tracker is calculated on a lower cost of calculation and full trail prediction and tracking on the real-time [21].

2. Framework

Before you begin to format your paper, first write and save Fig. 1 represents the overall architecture of proposed model. For simplicity, a brief description of overall model is given in this section.

Faster Region based CNN model is used to locate potential object in an image and classify them, respectively. The main aim of Faster R-CNN is to locate potential object and various background semantics. This is an important technique for locating objects, as locating potential object in multiple object scenes is confusing and difficult. Therefore, the object proposal network will use potential regions where the probability of finding an object is much higher. A naïve bayes algorithm is then used to develop correlation between object and background semantics.

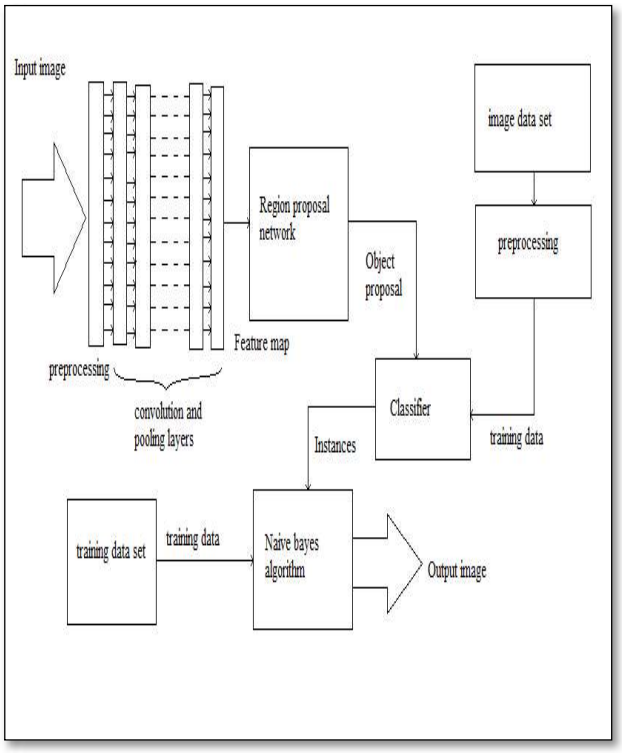

Figure 1 Block Diagram of Overall Model

A. Faster RCNN

Faster Region based CNN model plays a significant role in overall scene understanding process, because as it decreases the region of identification and increase the rate of Classification. A faster R-CNN-based Object Detection model is developed in this part to precisely locate potential object and find background semantics. As mentioned earlier, we ought to develop a scene understanding technique for a given scenario which is full of ambiguity. As a consequence, the learning pattern such as transfer learning paradigm can't be accessed [8]. In which the CNN model has unique representation features, this following feature layers provide general facilitation extraction capabilities. Therefore, we develop a region based method for locating potential objects.

R-CNN is the regions based semantic segmentation technique build on object detection out turns. Peculiarly R- CNN uses discriminating properties to extracts immense amount of object occurrence probability regions and then determines CNN features for them. Region proposal network in faster R-CNN is used for locating potential regions and latter this location is used by selective search to generate region proposals for detecting objects. The out turns of the regions based approach which predicts the probability of an anchor to being as a potential object or background framework and then concrete the anchor. We also acquire a helpful labeling method to help labelling method to help people identify large numbers of raw data and boost the practicality of supervised training. The faster R-CNN algorithm and naïve bayes algorithm are presented in Algorithm 1.

B. *Region Proposal Network*

Region proposal network utilizes the anchors for discovering various different regions for an object. Objectness score is used to predict the probability of an anchor is an object. Then a sliding window runs locally on this feature map. Sliding window size is n × n (3 × 3 here). For each sliding window a set of 9 anchors is created, in which everyone has the same center (xa, ya) but there are 3 different aspect ratios and 3 different scales. Note that all these indices are calculated in relation to the original image. The faster R-CNN model is having a special type of anchor generator with 16 x 16 width and height stride and it uses L2 regularization. The input of faster R-CNN is a low resolution frame representing potential objects in various regions with class labels. The main aim of faster R-CNN is to locate potential object in a specific region, and its extract boundaries and orientation of the object for better accuracy. The second thing is that if you want to reuse the trained network like CNN in the process then there is a noticeable area. Ensure that the receptive field of each location represents all anchors on the map. The region proposal network (RPN) in faster region based convolutional neural networks (Faster R-CNN) is used to reduce the computing requirements of the overall estimation process. RPN scans each location faster and efficiently in order to assess the need to further the process given region.
It does that whether or not there are 2 scores representing the probability of an object on each score by outputting the bounding box proposals.

C. *Anchor*

For every sliding-window location, we predict many state proposals at the same time, where each space is shown as the number of possible proposals. Therefore, the layer encodes 4K output coordinates, and the CLS layer shows 2k scores, which do not predict the probability of the object or the object for each offer 4. The parameter is related to the reference box, which we call anchor. An anchor is focused on the sliding window of the question, and is connected to the scale and the aspect ratio. By default, you use 3 measures and 3 aspect ratios which get K = 9 anchors in each sliding position. If you follow the anchor labeling process, you can also remove an anchor based on the same criteria as the resistant to refine. Here's an issue that the anchor labeled as a background should not be included in the regression because they do not have the right box on the ground.
Anchor boxes are just references, to accommodate different types of objects, they are selected to keep different aspect ratios and scales, for example, long objects like buses did not display correctly by Square Bounding Box. At each location of the convolutional layer, the bounding box regression head outputs the bounding box offsets for each anchor box while the classification layer outputs the objectness score, indicating whether the object exists or not, not every anchor box. Only those boxes are processed further with the same high likelihood of the object.

$$p^* = \begin{cases} 1 \;\; if \; IoU > 0.7 \\ -1 \;\; if \; IoU < 0.3 \\ 0 \;\; otherwise \end{cases} \quad (1)$$

$$IoU = \frac{anchor \cap GTbox}{anchor \; U \; GTbox} \quad (2)$$

$$L_{LOC}(t, v) = \sum_{i \in \{x,y,w,h\}} smooth_{L_1}(t_i^u - v_i), \quad (3)$$

$$smooth_{L_1}(x) = \begin{cases} 0.5\, x^2 \;\; if \; |x| < 1 \\ |x| - 0.5 \;\; otherwise \end{cases} \quad (4)$$

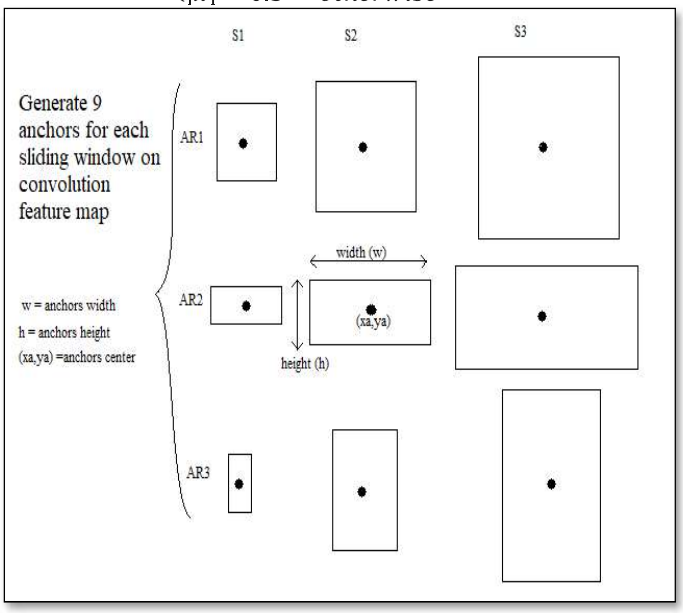

Figure 2 Anchor Generator Diagram

D. *Region of Interest*

After RPN, we get different sizes in different areas of the proposed areas along with CNN feature maps. An efficient design is not easy to work with different size features. The interest area makes it easy to reduce problems by reducing the feature maps to the same size. Contrary to Max-pooling at a fixed size, ROI pooling can make a certain number of calls to the input field and apply maximum pooling to almost all areas. This is why the ROI pooling product is not always considered to be the input size. The main objective of such aggregation is to increase the time of training and testing and end-to-end the entire system. It's a kind of pooling layer that performs

maximum pooling on non-uniform size inputs and a small size map of a fixed size. The choice of this fixed size is that the network is a hyper-parameter and is predefined. Fast R-CNN has introduced ROI level and is a special type of special pyramidal pooling layer, which is known for its visual detection in Deep Conventional Networks in Spatial Pyramid Pooling. The main function of the ROI level is to renew the figures with an arbitrary size in the size of a particular length due to size objection from fully connected levels.

E. Naïve's Bayes Algorithm

The Naive Bayes is based on the classifier so-called Bayesian theorem and is particularly useful when the input has a high dimension. Regardless of its simplicity, Naive Bayes can often out form more advanced classification task. Whether the Naive Bayes classifier can independently control the number of different variables, is constant or categorical. Even though predictions are considered independent, classification cannot be dramatic, because each character class can be used to do mathematics independently from conditional density, which means that it has many multi-tasking functions with reduced Dimensions. Naive Bayes reduces the work of estimating the high-dimensional density of one-dimensional kernel density. In contrast, perceptions do not affect old chances, especially in decision-making areas; therefore classification work is not affected.

$$p(c_j|x_{1,}x_2\ldots\ldots,x_d) \propto p(x_1,x_2 \ldots\ldots,x_d|c_j)p(c_j) \quad (5)$$

$$p(x|c_j) \propto \prod_{k=1}^{d} p(x_k|c_j) \quad (6)$$

$$p(c_j|x) \propto p(c_j)\prod_{k=1}^{d} p(x_k|c_j) \quad (7)$$

Where $p\left(c_j\middle|x_{1,}\ x_2 \ldots x_d\right)$ is posterior probability

$C = \{c_1, c_2 \ldots c_3)$Is possible outcomes and

$X = (x_1, x_2 \ldots x_d)$ Is set of variables

The naive's bayes classifier measures the probability for each element. Then it chooses the result with the highest probability. They are potential, which means they use the Possibility of every element for the given text, and then output the element with the maximum value. it describes the possibility of a features based on the use of bayes theorem, which is based on prior knowledge of the conditions related to that features.

3. Experimental Results

In this field, we have shown experimental results and analyze them. We use GPU and python language with deep learning tensorflow library as a platform. For training the model it requires a huge time as it take 24 to 50 sec per step. Training of the model is smooth and has quit good accuracy as it revolves in between 0 and 2. Certainly there is a huge fluctuation in the training graph as we have taken a wide variety of image pixel resolution which consolidates system in general but it increases training time as well as fluctuation in training measures. We were training 6 classes.

4. Discussion

If we analysis the objectness loss function, It shots beyond 0.160 and exponentially decreases with fluctuations to certain point and tends to saturate beyond certain values. Classification loss have a decent view as it has exponential nature and loss decreases as training increases, Classification accuracy based on this measure gives us a better performance of our model. We have classification accuracy ranging from 70% to 96%. if we visualize localization loss for a given object, model trends to have difficulty in locating an object and the loss is max which is nearby .01 to .2. But as the training proceeds the rate decreases exponentially and is nearly about proceeding to zero. Queue has an exponential start but after a certain point the queue get saturated and having constant values. Clone loss is similar if you are trained on multiple GPUs: Tensorflow will train a model to train each GPU and report the loss on each clone. If you provide model training on a GPU / CPU, then you can only see one clone loss, which is similar to total loss.Fig.3. Show a schematic image of our result which is taken at a step of 1534. Detection accuracy is lower as it detected only four objects with pretty good accuracy ranging from 70% to 96%. Under such detection circumstances the final result of our model is pretty excellent which label scene as classroom on the basis of four object detected in an image.

Algorithm 1

Input :- Data frame I (from Faster RCNN), scene template O
Output:- labelled scene L

Scene_list = *null;*

for *i* ∈ *I* ***do***, */* searching valid data in data frame*/*
for *o* ∈ *O* ***do***, */*trying all scene templetes */*

if is_scene_labeling ***then***,
/ execute labelling to a scene*/*

task(i) = P(feature = instance/class =task) p(class=task);*

if *task(i) >threshold* ***then***
append *(i)*to scene_list.

else */* import to RCNN model*/*
task(i), locate(i) = faster RCNN ←i;

if *task(i) > threshold* ***then***,
append locate*(i)* to scene_list.

return scene_list;

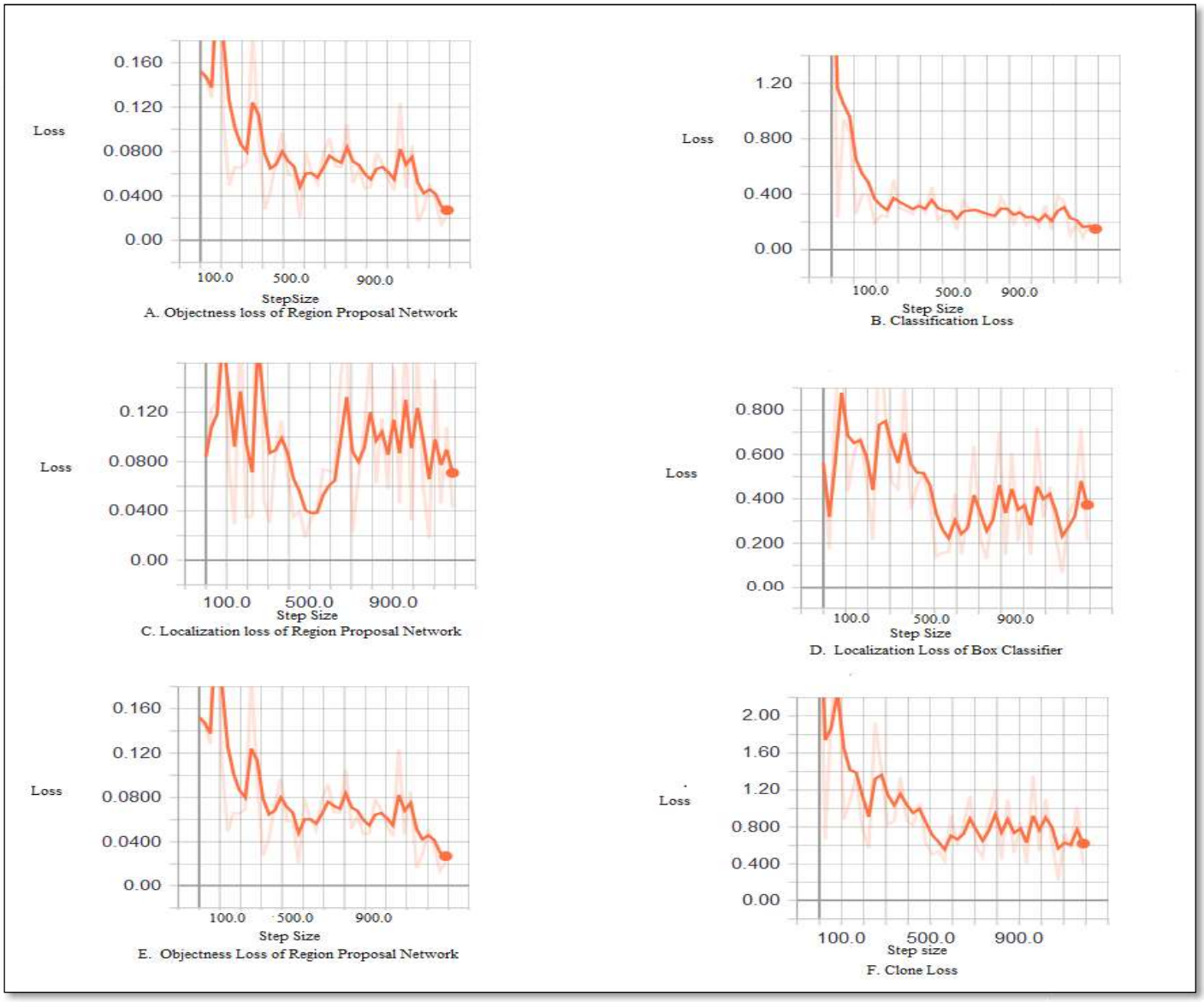

Figure 3 Loss Function

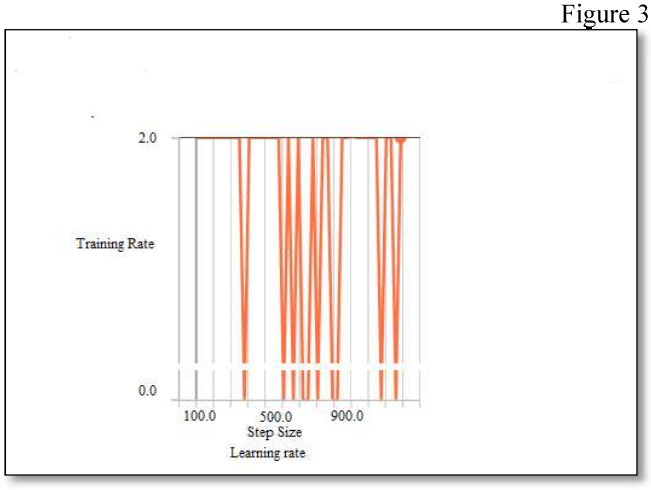

Figure 4 Learning Rate

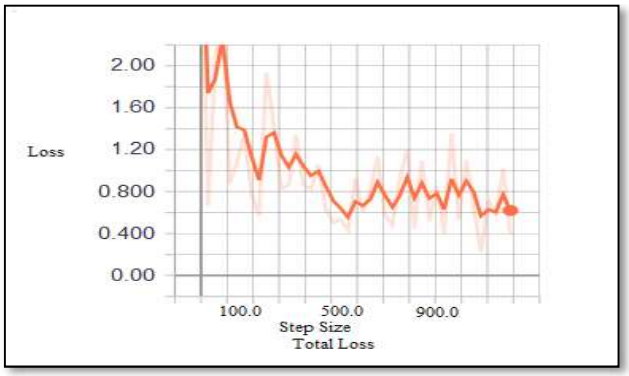

Figure 5 Total Loss

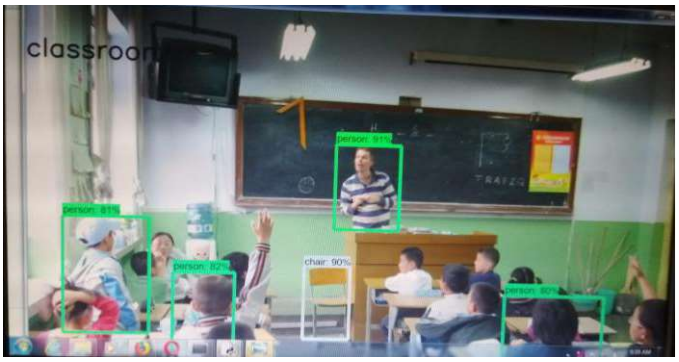

Figure 6 Output Image

5. Conclusion

A faster R-CNN-based scene understanding method is introduced in the paper. The most knowledgeable benefits of this method are RPN to make efficient and accurate field proposals, sharing interactive features with down-stream detection networks, our proposed method enables integrated, intensive-study-based scene understanding systems to run near the actual real time frame rate. The learned scene understanding model also improves the quality of the province and thus identifies the overall scene.

References

1. Shao, J., Joy, C. C., Kang, K., & Wang, X. (2016). "Crowded scene understanding by deeply learned volumetric slices". *IEEE transactions on circuits and systems for video technology*, 1-11.
2. Wojek, C.,Rroth, S., Schindler, K., & Schiele, B. (april 2013). "Monocular visual scene understanding: understanding multi-object traffic scenes*". IEEE transactions on pattern analysis and machine intelligence*, 882-897.
3. Ren, S., He, K., Girshick, R., & Sun, J. (2016). "Faster r-cnn: towards real-time object detection with region proposal networks*". IEEE transactions on pattern analysis and machine intelligence*, 1-14.
4. Li, H., Huang, Y., & Zhang, A. Z. (august 8, 2017.)." An improved faster R-CNN for same object retrieval". 13665-13676.
5. Chen, X., Kundu, K., Zhu, Y., Ma, H., Fidler, S., & Urtasun, R. (n.d.). "3d object proposals using stereo imagery for accurate object class detection". *IEEE transactions on pattern analysis and machine intelligence*, 1-14.
6. Li, J., Won, H.-C., Lo, L.-L., & Xin, Y. (february 2018). Multiple object detection by a deformable part-based model and an R-CNN*". IEEE signal processing letter,*288-292
7. Akcay, S., Kundegorski, M. E., Willcocks, C. G., & Breckon, T. P. (2018). "Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery*". IEEE transactions on information forensics and security*, 1-13.
8. Taheri, S., & Toygar, O. (march 2018). "Animal classification using facial images with score-level fusion". *IET computer vision* , 679-685.
9. Szatmári, I., Schultz, A., Rekeczky, C., Kozek, T., Roska, T., & Chua, L. O. (september 2000). "Morphology and autowave metric on cnn applied to bubble-debris classification". *IEEE transactions on neural network*, 1385-1393.
10. Sahbi, H., Audibert, J.-Y., & Keriven, R. (april 2011). "Context-dependent kernels for object classification*". IEEE transactions on pattern analysis and machine intelligence* , 699-708.
11. Zhang, L., Zhen, X., & Shao, L. (august 2014). "Learning object-to-class kernels for scene classification". *IEEE transactions on image processing,* 3241-3253.
12. Zhang, Z., Zhao, Y., Wang, Y., Liu, J., Yao, Z., & Tang, J. (october2013). "Transferring training instances for convenient cross-view object classification in surveillance". *IEEE transactions on information forensics and security*, 1632-1641.
13. Park, S.-W., Park, J., Bong, K., Shin, D., Lee, J., Choi, S., et al. (2015). An energy-efficient and scalable deep learning/inference processor with tetra-parallel mimd architecture for big data applications. *IEEE transactions on biomedical circuits and systems*, 1-11.
14. Zhang, T., Zheng, W., Cui, Z., Zong, Y., Yan, J., & Yan, K. (2016). "A deep neural network driven feature learning method for multi-view facial expression recognition". *IEEE transactions on multimedia* , 1-10.
15. Fan, J., Zhao, T., Kuang, Z., Zheng, Y., Zhang, J., Yu, J., et al. (2016). "Hd-mtl: hierarchical deep multi-task learning for large-scale visual recognition". *IEEE transactions on image processing,* 1-17.
16. Di, S., Zhang, H., Li, C.-G., Mei, X., Prokhorov, D., & Ling, H. (2017). "Cross-domain traffic scene understanding: a dense correspondence-based transfer learning approach". *IEEE transactions on intelligent transportation systems* , 1-13.
17. Kurdej, M., Moras, J., & Bonnifait, V. C. (2013). "Map-aided evidential grids for driving scene understanding". *IEEE intelligent transportation systems magazine*, 30-41.
18. Li, L., Ota, K., Dong, M., & Borjigin, W. (2017). "Eyes in the dark: distributed scene understanding for disaster management*". IEEE transactions on parallel and distributed systems* , 1-14.
19. Lim, C. H., Risnumawan, A., & Chan, C. S. (2013). 'Scene image is non-mutually exclusive a fuzzy qualitative scene understanding". 1-16.
20. Liu, X., Yang, W., Lin, L., Wang, Q., Cai, Z., & Lai, J. (2015). "Data-driven scene understanding with adaptively retrieved exemplars". *IEEE computer society,* 82-92.
21. Liu, Y., Wang, Q., Zhuang, Y., & Hu, H. (october 15, 2017). "A novel trail detection and scene understanding framework for a quadrotor uav with monocular vision". *IEEE sensors journal*, 6778-6787.