ZF Model Efficiency for Automatic Hand Detection in Vehicles

1st Sofiane Medjram
LISTIC-Lab, Savoie Mont Blanc University
Polytech Annecy Chambéry
Annecy, France
medjram.sofiane@gmail.com

2nd Saida Bouakaz
Department of Computing Science
Claude Bernard Lyon 1 University
CNRS LIRIS UMR5205 F-69622 Lyon, France
saida.bouakaz @univ-lyon1.fr

***Abstract*—Automatic Hand detection in vehicles achieved best performances using deep learning approach. MS-FRCNN and MS-RFCN are the models that achieve the state of the art results on the viva hand detection challenge. These models, unfortunately, require important resources to operate (more than 7 Giga bytes GPU) which presents an important limit for their large scale applications. In this paper, we address the viva challenge of hand detection in vehicles using simpler models and low computer components (2 Giga bytes GPU). We found that the ZF model is highly adequate for the problem and more importantly, it overpasses remarkably the precisions of complex model of literature MS- FRCNN. The experiment result of ZF model on the viva database shows (AP = 91.5%, AR = 85.7% at level one and AP = 80.5%, AR = 69.5% and Fps = 0.4 s at level two in measures of Average Precision (AP), Average Recall(AR) and Frame per seconds (Fps). These results of ZF reached the fourth place on the viva hand detection challenge. Furthermore, relying on the lateral inhibition concept that allow better generalization for non-learned cases, we enhanced the ZF model precisions by its application on both modules (RPN , RCNN). The application reduces the false positives results presented in the ZF model and the precisions obtained are less than the ZF model and better than many models of literature including Yolo, Modified-Faster Rcnn and ACF.**

***Index Terms*—Automatic hand detection, deep learning, ZF model, lateral inhibition, normalization layer, Faster R-CNN**

I. Introduction

The hand detection problem has been the subject of interest of recent work of literature, it has been studied in several application domains including the new generation of smart vehicles. In fact, with the appearance of deep learning and convolution neuronal networks (vision-based), the hand detection achieves the best performances in front of most difficult challenges including hand variations, occlusions, low-resolution, lighting changes, varied hand shape and viewpoints [1][2][3]. The secret behind the best performances is due to the parameters sharing and sparsity connections proprieties of convolution networks. They allow a robust decomposition and accumulation of feature descriptors for hand regions. The methods of hand detection in vehicles using deep learning [1][2] have adapted the models of Faster R-CNN framework [4], VGG16 [5] and Resnet [6] respectively. These models require lots of resources to operate (7 giga bytes GPU memory for VGG16 and 8 giga bytes GPU memory for Resnet) which create an important problem for their application on computer low components-based. In this paper, we exploited and adapted the problem of hand detection in vehicles using the lightest version of Faster R-CNN : the model of ZF (Zegler and Fergus) [4]. The experimentation results showed that ZF model handle remarkably the problem of hand detection in vehicle using low computer components and achieved the fourth place in the viva hand detection challenge. The ZF model is shown in Figure 1. The model is learned using the Caffe Framework [7] and evaluated using the Viva Hand Detection Challenge Database [8]. The proposed method achieves remarkable precisions and speed in front of the challenges presented in the viva database (L1-AP = 91.5%, L1-AR = 85.7% / L2-AP = 80.5%, L2-AR = 69.5% and Fps = 0.4 s) (see Figure 2). The rest of paper is organized as follows. In Section2, we present some works of hand detection in literature, we present the standard architecture of Faster R-CNN using ZF, VGG and Resnet models and finally, we present a brief presentation of the viva hand detection database used in our experiment. In Section 3, we explain how we adapt the ZF model to the problem of hand detection. Next, in Section 4, we illustrate how we proceed to learn the ZF model with the obtained performances in measure of AP, AR and Fps. Finally, in Section 5, we conclude the work.

Identify applicable funding agency here. If none, delete this.

II. Related work

The hand detection problem presents an active field of computer vision and it has attracted numerous application domains such as augmented reality [9] [10], virtual reality [11] [12], human computer interaction [13] [14] [15] [16] [17] and smart vehicles [1] [2] [18] [19]. In this section, we will interest to the methods designed for smart vehicles. Several methods have been proposed, in this study, we divided them in two categories: classic methods and deep learning methods.

A. Classic Methods

The classic methods of hand detection start by extracting features from the image beforehand then, they try to generalize them by applying some machine learning algorithms. Mittal et al [18] proposed a well performing automatic hand detection method based on multiple proposals. The method is based

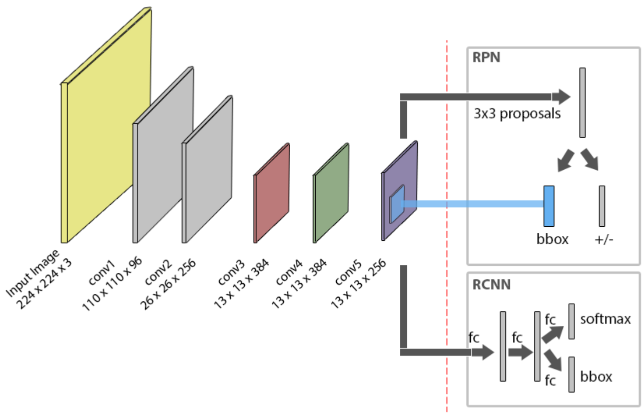

Fig. 1. The light version of Faster Rcnn Object Detection (ZF model : 8 layers over 2 connected modules RPN+ RCNN)

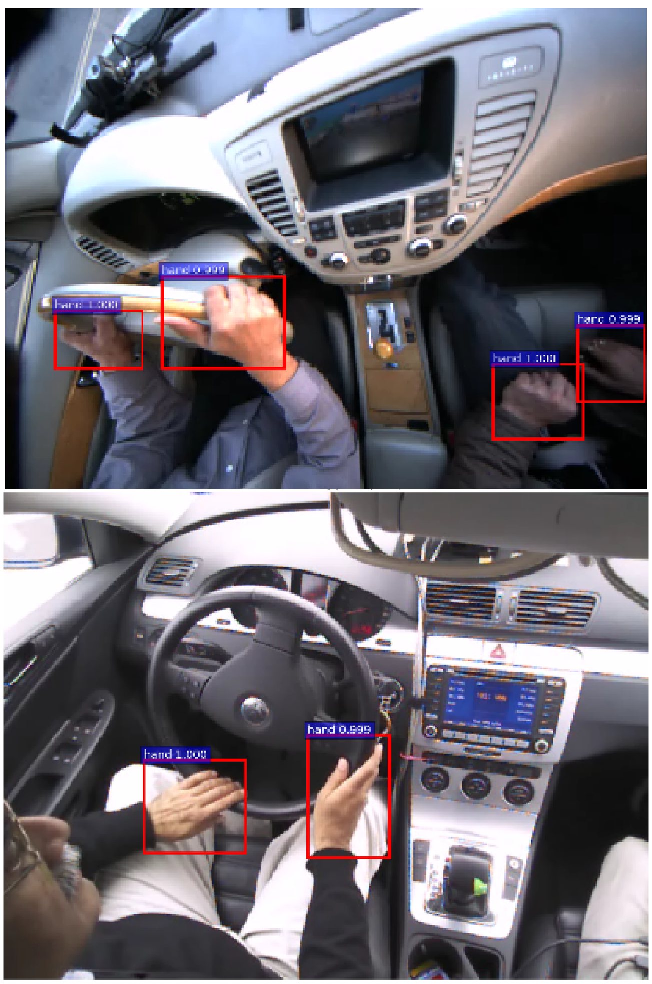

Fig. 2. Some results of hand detection obtained from the Viva hand database using the light version ZF models of Fatser R-CNN.

on two-stages: hypothesize bounding boxes and then, their classification. The hypothesize stage represents the results of three independent hand detectors (sliding window hand-shape, context-based and skin-based detectors), then the proposal regions are passed to a learning phase for their classification using a confidence score, the common results between these three proposals will be considered as hand regions. Zhou et al [19] handled the problem of hand detection using the concept of context aware hierarchy. Due to the sensitivity of classic descriptors in front of difficult challenges (strong lighting changes, occlusions, shadows...), the authors have incorporated the concept of context-aware in order to generalize well the detection of the hand. From the input image, they start by estimating the camera viewpoint, then a pre-proposals hand regions will be estimated. These last will be refined and cleaned using the information of context-aware (shape-fitting, spatial filtering and mutual voting). Although these methods succeeded to provide acceptable results of hand detection in vehicles, they are still sensitive to the difficult challenges appearing in the hand scenes, which make their applicability very restrictive and limited.

B. Deep Learning Methods

The deep learning methods are based on a deep neuronal model (more than one hidden layer) accompanied with a big set of examples (more than 3000 images with their annotations), then through backpropagation iterations the model learn how to detect the object. Several frameworks have been developed for object detection where the hand detection problem makes part, the most known frameworks are Faster R-CNN [4], Yolo [20] and SSD [21]. Faster R-CNN is the most used for hand detection. The hand detection problem for viva challenge has been exploited in the literature using two models of Faster R-CNN, VGG and Resnet. The pre-trained value available for these models have been trained and evaluated on the famous databases Coco [22] and Pascal Voc [23] to detect and to localize several classes in the image (about 20 classes). According to the authors of [1] [2], the ROIs (region of interests) function of hands are much complex than the ROIs function of Coco and Pascal Voc classes. The performances of the application of VGG and Resnet models (using the pre-trained values of Coco and Pascal Voc) for the hand viva challenge are less robust, the models do not generalize well for the case of hand images. Therefore, to overcome this lacuna, the authors modified the models by adding some extra layers with normalization to achieve the performance needed/wished. In [1], the authors modified the model of VGG16 to produce a new model named MS-FRCNN (Multiple Scale Faster Region-Based Convolutional Neural Network) (see Figure 3). Based on their experiments, they argued that the VGG16 is not suitable for hand detection. Due the small region spaces that the hand occupies in the image, due to their low resolutions as well as the difficult challenges presented in the scene (lighting changing, occlusions..), the features maps representing the spatial information of the hands are less representative. In addition, because of the small regions of the hand, the information collected along the deeper layers of VGG16 are collected outside the region of interest, which is not significant. Therefore, to resolve these problems, the authors combined the local features (last layer) with the global features (some previous layers) to enhance the description quality of the hand, and that in both stages (RPN an RCNN). In addition, to avoid that the values of a specific layer in the combination phase dominate, a normalization operation is applied to every layer in order to equilibrate their contribution in the combination.

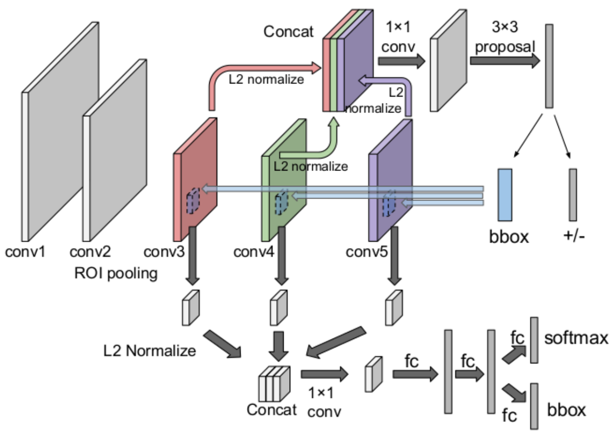

Fig. 3. MS-FRCNN model proposed in [1] for hand detection in vehicule (modified version of VGG16 Faster R-CNN)

In [2], the authors enhanced their previous work using a deeper model, Resnet with 101 hidden layers instead of 16 of VGG. In such model, the features maps generated at the last layer will be more representative but very sensitive to small scale regions. However, to enhance the capability of Resnet in front of small scale regions of the hand, a concatenation of several features maps of previous shallower layers with the last layer is established, the final model produced is named MS-FCN (Multiple Scale Region-Based Fully Convolutional Networks) (see Figure 4). In MS-FCN, low-level localization information (last layer) with high-level semantic information (some previous layers) are regrouped. In fact, the main difference between the MS FRCNN and MS FCN is that the first one is designed for hand detection in vehicles only, the second one instead is for hand detection in vehicles and the wild. Although these modified Faster R-CNN models presented the state of the art results on the viva hand detection challenge, they consume lot of resources to operate which present their major limit.

After mentioning the most recent methods of hand detection in vehicles using deep learning approach, it will be convenient to present the framework that they have been based on : Faster R-CNN.

C. Faster R-CNN

Fatser R-CNN [4] is one of the most famous neuronal network applied for object detection. It is composed of two modules, the first module RPN (Region Proposals Network) starts by hypothesis region proposals then the second module Fast R-CNN (Fast Regional Convolution Neuronal Network)

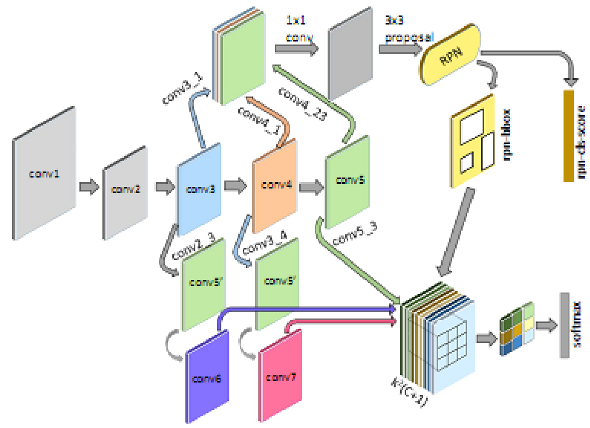

Fig. 4. MS-FCN model proposed in [2] for hand detection in vehicule (modified version of Resnet Faster R-CNN)

tries to classify the proposed regions (see Figure 5). These modules RPN and Fast R-CNN have been unified in one single network through the sharing property of the convolution operation. In fact, several version of Faster R-CNN have been proposed, we cite ZF (8 layers), VGG (16 layers) and ResNet (101 layers) models. These versions are different in term of the total number of parameters learned in both modules, the number of layers as well as the number of shared layers. These differences will affect the performances of the detector for a particular problem, generally the more complicated or deeper network is, the more time and resources it will consumes and the more precision it will gain.

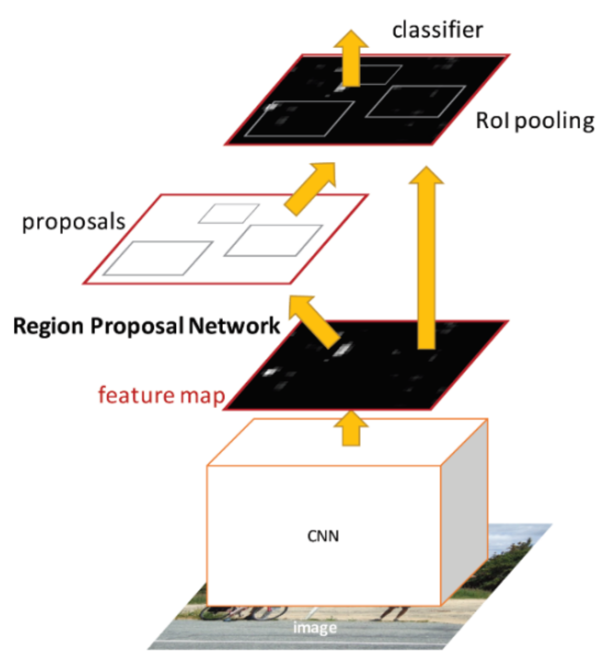

Fig. 5. Faster R-CNN architecture flowchart describing the connectivity and the sharebility of RPN and Rcnn modules.

D. Viva Hand Detection Challenge

The VIVA[8] (Vision for Intelligent Vehicles and Applications) challenge presents a big number of hand images in vehicles designed for machine / deep learning applications. It contains 5500 images for learning and 5500 for testing, the learning images are accompanied with hands annotations reflecting their positions in the images and the testing images are for generalization. The images are collected from 54 videos all over the web, captured from 7 viewpoints under uncontrolled environments, including illuminations changing and occlusions

III. Proposed Method

The success of any object detection network in the deep learning approach is based on its efficiency in region proposals stage and/or their refinement in the classification stage. Using the Faster-RCNN framework and based on the concept of starting learning with the simplest model first [24], we chosen the ZF model. As we mentioned above, the ZF model represents the light version of Faster-RCNN, we could run the learning phase on a computer with low components (2 Gigabytes GPU). For which concern the difficult challenges presented in the viva dataset and how they have been treated by the authors in [1] [2], we considered them differently (using the ZF model). Theoretically, in the context of deep learning, the difficult challenges presented in the viva dataset as (low resolution, strong lighting changing, occlusions) should be treated implicitly with the model trough the decomposition over layers and deeper volumes of feature maps. The suitable model you have with the sufficient data you provide will define adequate parameters for these challenges. Therefore, our experiment using the ZF model on the database of Viva hand detection shown that the model succeeded to detect the majority of the hand regions on the test set images (L1-AP = 91.5%, L1-AR = 85.7% / L2-AP = 80.5%, L2-AR = 69.5% and Fps = 0.4 s) with some false positives in some viewpoints of camera. The first experiment has confirmed the theory mentioned above, the accumulation of parameters over deeper layers in the ZF model resolved the problems of difficult challenges easily. For which concern the false positives detection resulted by the ZF model, it affirmed that the function complexity of the network shares some properties of the object of interest with other objects of the scene. The ZF model presents an over-fitting problem in regions proposals stage and/or their classification. Compared to the previous models MS-FRCNN and MS-FCN used in [1][2] respectively, these last are much deeper and much complex than the ZF model (8 layers). As their authors mentioned on their papers, the VGG16 and Resnet models of Faster R-CNN are not suitable for hand detection, especially for the challenges accompanied. Therefore, they combined the parameters of previous layers with the last one in both modules RPN and RCNN. The ZF model, in contrast, showed (independently of the false positive results detection) that it is sufficient to detect well the hand regions in the scene, we can consider the ZF model as the normalized version of VGG16 and Resnet were lots of layers have been removed and because of that the ZF has adapted well with the problem of small regions that the hands occupies in the images. The deep learning approach is mainly based on empirical sequences of experimentation. Inspiring from the previous experiments exploited in [1][2], we interested to apply the concepts of layers concatenation and/or normalization on the ZF model in order to know if they will reduce or not the amount of false positive detections. However, through our experiments we found that there is no need for layers concatenation, the precision obtained by the feature maps tensors of ZF model describe the Regions of interests of the hands much better than the concatenation models. Based on this result, we continued our refinement on the false positive results using the concept of lateral inhibition (Normalization), this last serves as aids for generalization as it has been mentioned by their author in [25]. The hands are commonly presented on images as small regions and many object of the scene can share some appearances properties of it. Using the ZF model over its deeper layers the features of the hands learned become less representative and will generate lots of false detection. We applied the concept of lateral inhibition on the final features maps tensors (conv5) of both modules and we found that by encouraging the participation of neighbor pixels of feature maps tensors, the resulted models reduced the amount of false detection but lose some precision in localization. Although the normalized models of ZF lose some precision in the localization (see table 1, Figure 10), they present better results than the ZF model in some viewpoints of camera and better results than some models of literature, which made them recommended for some particular cases. In the following, we presents the architecture of the three normalized models of ZF with the corresponding formula of lateral inhibition.

$$b^i_{(x,y)} = x^i_{(x,y)}/(k + \sum_{j=max(0,i-n/2)}^{min(N-1,i+n/2)} (a^i_{(x,y)})^2)^b \qquad (1)$$

where

ai(x, y) represents the i th conv. kernel's output (after ReLU) at the position of (x, y) in the feature map.

bi(x, y) represents the output of local response normalization, and the input for the next layer.

N is the number of the conv. kernel number.

n is the adjacent conv. kernel number.

k, , are hyper parameters.

IV. Experiment and Results

As we mentioned above, we start our experiment using the simplest model of Faster R-CNN : the ZF model, and from its results a successions of refinements and modifications have been applied in order to know if the results of the model can be enhanced. Our experiments are conducted on Hp Z-book, Nvidia Quadro K2100m 2G, using the database of viva hand detection which contains 5500 train images of hands. For the evaluation, we divide it in two stages. In the first stage, we compared the results of the ZF model with the results of

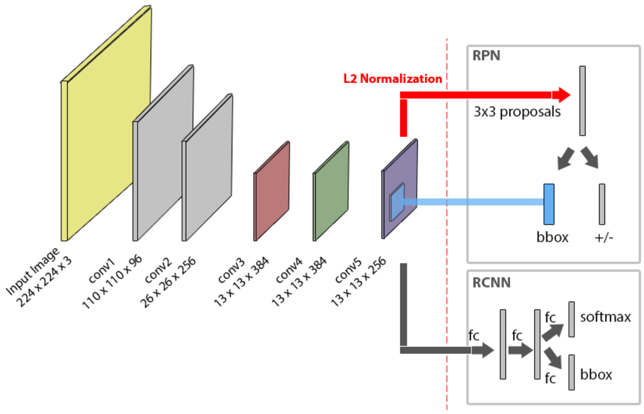

Fig. 6. Normalized RCNN ZF model.

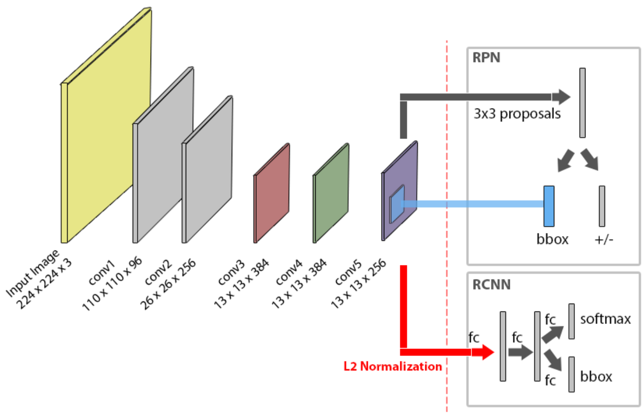

Fig. 7. Normalized RPN ZF model.

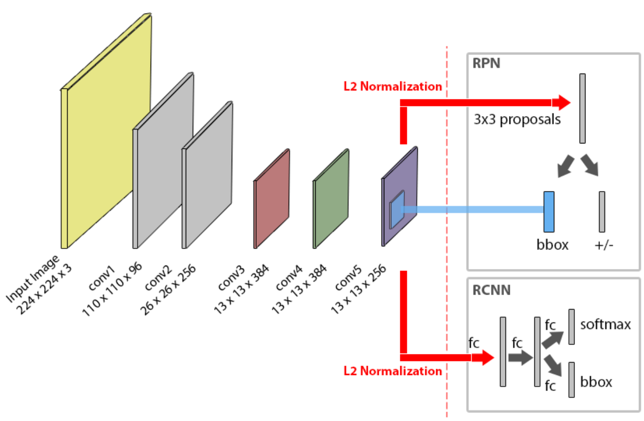

Fig. 8. Normalized (RPN + RCNN) ZF model.

best models of literature MS-FRCNN [1] and MS-FCN [2] in measure of average precision (AP), average recall (AR), time speed (Fps), number of layers and GPU memory requirement. In the second stage, we compared the ZF model with the results of the three normalized models explained above and that in measure of average precision (AP), average recall (AR) and time speed (Fps). Therefore, in order to choose the final model for the problem addressed, we respected the standard recipe of any project of deep learning which is dividing the dataset in training set and dev set. We divided 99% of training images provided by viva as train set and the remaining 1% as dev set. Next, we learned the ZF Model of Faster R-CNN on the training set over 10000 iterations using the pre-trained weights of Pascal VOC dataset [23] and we set the learning rate base to 0.001. The average precision of the ZF model on the dev set was 99.5% and its generalization for the test set was as follow (L1-AP = 91.5%, L1-AR = 85.7% / L2-AP = 80.5%, L2-AR = 69.5% and Fps = 0.4 s). Compared to the best methods of literature, the ZF model succeeded to overpass the performance of MS-FRCNN model and that in measure of precision, recall, number of layers as well as GPU memory requirement. The following, Table 1 Figure 9 and Figure 10, describe and compare the performances of ZF, MS-FRCNN and MS-FCN models obtained using the Viva hand detection database.

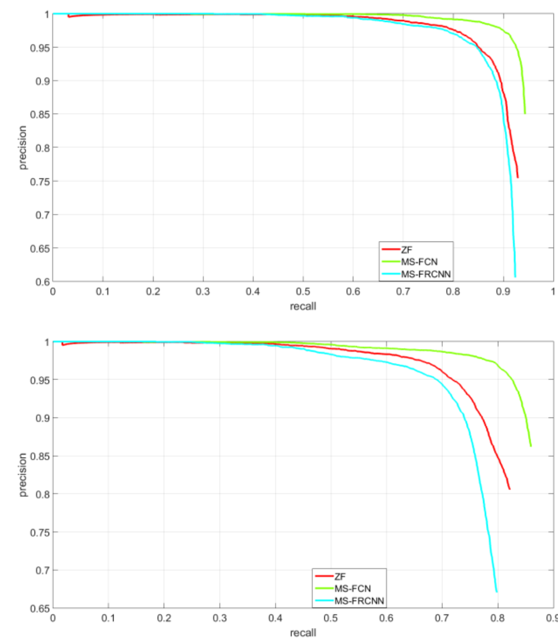

Fig. 9. Roc Curves, AP (average precision) and AR (average recall) obtained by [1] (green) , [2] (blue) and ZF (red) models on the Viva database. (up) L1-AP, (down) L2-AP.

The analysis of the ZF model precision showed that the model does not generalize well on image tests of viva challenge, some false positive results are presented in some viewpoints of camera (see Figure 11). Technically, we interpret the false positive results of the ZF model as an over-fitting. In order to overcome this limit without the need of additional images with labels, we modified the neuronal network structure and we applied the regularization concept (Normalization). And because of Faster R-CNN framework is composed of two modules and the performance of the second module highly depend on the results of the first one, we modified the ZF model with respect to the three following options: Normalizing

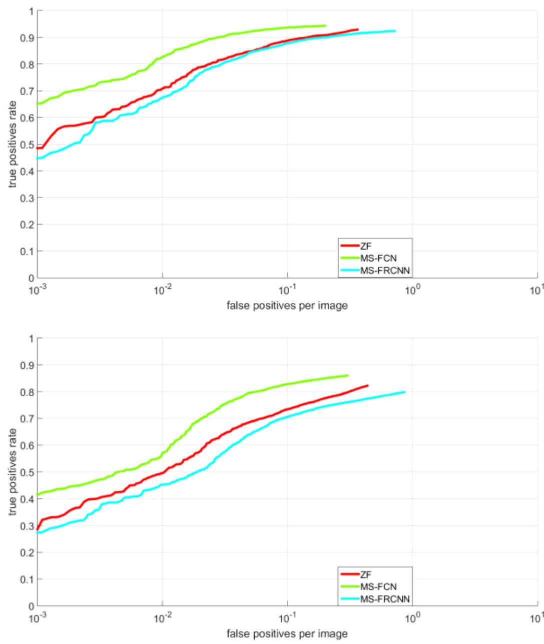

Fig. 10. Roc Curves, AP (average precision) and AR (average recall) obtained by [1] (green) , [2] (blue) and ZF (red) models on the Viva database. (up) L1-AR, (down) L2AR.

RPN only, Normalizing Rcnn only and Normalizing Both Modules together (RPN + RC NN) (see Table 2).

Table 1: Summary of different precision results obtained for models using in the experiments (ZF, MS-FRCNN, MS-FCN).

Model	EvalSets	Ap	AR	Fps	Gpu
ZF	L1 L2	91.5% 80.5%	85.7% 69.5%	0.4s	2g
MS- FRCNN 8g	L1 L2 MS- FCN	90.8% L1	84.1% 95.1%	/ 94.5%	7g 0.21s%
 MS- FCN	L2 L1 L2 77.6%	dfsd 95.1% dfsd 65.1%	sdfsd 94.5% sdfsd	0.21s%	8g
MS- FCN	L1 L2	95.1% 86.0%	94.5% 83.4%	0.21s%	8g

Table 2: Summary of different precision results obtained for normalized version of ZF model.

	RPN+L2	Rcnn+L2	ES	AP	AR	Fps
ZF	no	no	L1 L2	91.5% 80.5%	85.7% 69.5%	0.4
NZF1	yes	no	L1 L2	83.6% 68.8%	81.1% 59.9%	0.36
NZF2	no	yes	L1 L2	78.8% 63.8%	72.8% 53.4%	0.4
NZF3	yes	yes	L1 L2	86.9% 72.8%	81.1% 61.8%	0.42

CONCLUSION

In this paper, we showed how the Faster RCNN ZF model adapt adequately with the problem of hand detection in vehicles without the need of complex models and important computer resources, such as MS FRCNN or MS FCN. We showed that the performances obtained by the ZF model surpassed many models of literature and reached the fourth place in the viva hand detection challenge. In addition, we found that the Faster R-CNN models (ZF, VGG, ResNet) precisions highly depend on the problem addressed. The manner that the problem is defined changes from one model to another. In other words, using a complex model, the problem of the hand detection becomes much difficult and complex (as mentioned in [2]), but using simpler model, like ZF model, the problem is simplified. To avoid this confusion, it is convenient to start learning with simplest model first, always. We showed also that even with the satisfying precision that the ZF model presents, lots of false positives results are presented in some viewpoints. By applying the normalization concept, we succeeded to reduce this amount of false positives but we lost some precision of correct detection at the same time. As perspective, we work to find the best combination of hyper parameters values of the lateral inhibition that allows obtaining highest precision of detection with the lowest amount of false positives.

REFERENCES

[1] T. H. N. Le, K. G. Quach, C. Zhu, C. N. Duong, K. Luu, and M. Savvides, “Robust Hand Detection and Classification in Vehicles,” in 23rd International Conference on Pattern Recognition (ICPR) Cancún Center, Cancún, México, December 4-8, 2016, 2016, vol. December 4, pp. 573–578.

[2] T. H. N. Le, K. G. Quach, C. Zhu, C. N. Duong, K. Luu, and M. Savvides, “Robust Hand Detection and Classification in Vehicles and in the Wild,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2017, vol. 2017-July, pp. 1203–1210.

[3] N. Das, E. Ohn-Bar, and M. M. Trivedi, “On Performance Evaluation of Driver Hand Detection Algorithms: Challenges, Dataset, and Metrics,” IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC, vol. 2015-Octob, pp. 2953–2958, 2015.

[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017.

[5] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Int. Conf. Learn. Represent., pp. 1–14, 2015.

[6] S. Wu, S. Zhong, and Y. Liu, “Deep residual learning for image steganalysis,” Multimedia Tools and Applications, pp. 1–17, 2017.

[7] Y. Jia et al., “Caffe,” Proc. ACM Int. Conf. Multimed. - MM ’14, 2014.

[8] “Vision for Intelligent Vehicles and Applications (VIVA) Challenge.” [Online]. Available: http://cvrr.ucsd.edu/vivachallenge/. [Accessed: 08-Jul-2018].

[9] F. Abedan Kondori, S. Yousefi, J.-P. Kouma, L. Liu, and H. Li, “Direct hand pose estimation for immersive gestural interaction,” Pattern Recognit. Lett., vol. 66, pp. 91–99, Nov. 2015.

[10] F. Dominio, M. Donadeo, and P. Zanuttigh, “Combining multiple depth-based descriptors for hand gesture recognition,” Pattern Recognit. Lett., vol. 50, pp. 101–111, Dec. 2014.

[11] J. Choi, B. Seo, and J.-I. Park, “Robust Hand Detection for Augmented Reality Interface,” in Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry, 2009, vol. 1, no. 212, pp. 319–322.

[12] K. Pin Ng, G. Yew Tan, and Y. Ping Wong, "Vision-Based Hand Detection for Registration of Virtual Objects in Augmented Reality," Int. J. Futur. Comput. Commun., vol. 2, no. 5, pp. 423–427, 2013.
[13] C. Yang, D. K. Han, and H. Ko, "Continuous hand gesture recognition based on trajectory shape information," Pattern Recognit. Lett., vol. 99, pp. 39–47, Nov. 2017.
[14] G. Modanwal and K. Sarawadekar, "A robust wrist point detection algorithm using geometric features," Pattern Recognit. Lett., vol. 110, pp. 72–78, Jul. 2018.
[15] C. Oden, A. Ercil, and B. Buke, "Combining implicit polynomials and geometric features for hand recognition," Pattern Recognit. Lett., vol. 24, no. 13, pp. 2145–2152, Sep. 2003.
[16] S. Medjram, M. C. Babahenini, A. Taleb-Ahmed, and Y. Mohamed Ben Ali, "Real-time wrist localization in color images based on corner analysis," Multimed. Tools Appl., 2016.
[17] S. Medjram, M. C. Babahenini, A. Taleb-Ahmed, and Y. Mohamed Ben Ali, "Automatic Hand Detection in Color Images based on skin region verification," Multimed. Tools Appl., vol. 77, no. 11, pp. 13821–13851, Jun. 2018.
[18] A. Mittal, A. Zisserman, and P. Torr, "Hand detection using multiple proposals," Procedings Br. Mach. Vis. Conf. 2011, pp. 75.1-75.11, 2011.
[19] T. Zhou, P. J. Pillai, and V. G. Yalla, "Hierarchical context-Aware hand detection algorithm for naturalistic driving," IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC, pp. 1291–1297, 2016.
[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2015.
[21] W. Liu et al., "SSD: Single shot multibox detector," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016.
[22] T. Y. Lin et al., "Microsoft COCO: Common objects in context," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014.
[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," Int. J. Comput. Vis., 2010.
[24] M. D. Zeiler and R. Fergus, "ZFNet," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2014.
[25] A. Krizhevsky, I. Sutskever, and H. Geoffrey E., "ImageNet Classification with Deep Convolutional Neural Networks," Adv. Neural Inf. Process. Syst. 25, pp. 1–9, 2012.