Multiple Scale Faster-RCNN Approach to Driver's Cell-phone Usage and Hands on Steering Wheel Detection

T. Hoang Ngan Le, Yutong Zheng*, Chenchen Zhu*, Khoa Luu and Marios Savvides
CyLab Biometrics Center and the Department of Electrical and Computer Engineering,
Carnegie Mellon University, Pittsburgh, PA, USA
{thihoanl, yutongzh, chenchez, kluu}@andrew.cmu.edu, msavvid@ri.cmu.edu

Abstract

In this paper, we present an advanced deep learning based approach to automatically determine whether a driver is using a cell-phone as well as detect if his/her hands are on the steering wheel (i.e. counting the number of hands on the wheel). To robustly detect small objects such as hands, we propose Multiple Scale Faster-RCNN (MS-FRCNN) approach that uses a standard Region Proposal Network (RPN) generation and incorporates feature maps from shallower convolution feature maps, i.e. conv3 and conv4, for ROI pooling. In our driver distraction detection framework, we first make use of the proposed MS-FRCNN to detect individual objects, namely, a hand, a cell-phone, and a steering wheel. Then, the geometric information is extracted to determine if a cell-phone is being used or how many hands are on the wheel. The proposed approach is demonstrated and evaluated on the Vision for Intelligent Vehicles and Applications (VIVA) Challenge database and the challenging Strategic Highway Research Program (SHRP-2) face view videos that was acquired to monitor drivers under naturalistic driving conditions. The experimental results show that our method archives better performance than Faster R-CNN on both hands on wheel detection and cell-phone usage detection while remaining at similar testing cost. Compare to the state-of-the-art cell-phone usage detection, our approach obtains higher accuracy, is less time consuming and is independent to landmarking. The groundtruth database will be publicly available.

1. Introduction

According to a study released by the National Highway Traffic Safety Administration (NHTSA) and the Virginia Tech Transportation Institute (VTTI), 80% of car accidents involve driver distraction under different forms such as talking on a cell-phone, sending text messages, reading a book,

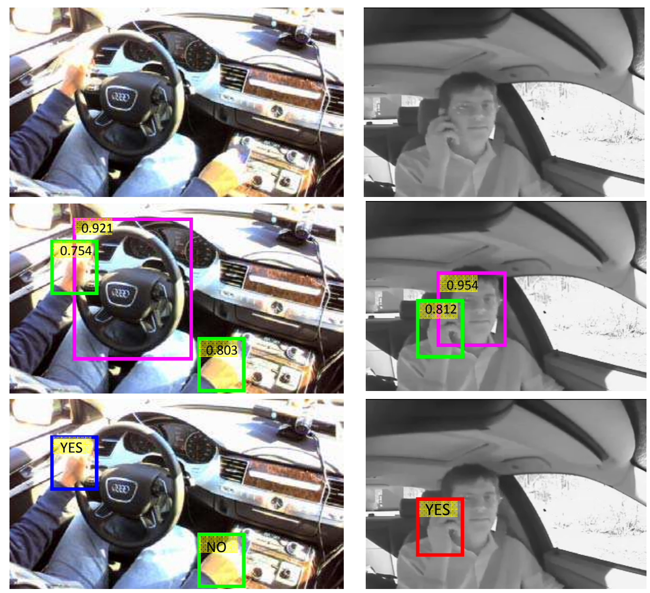

Figure 1. Our proposed Multiple Scale Faster-RCNN (MS-FRCNN) approach incorporating face detection and steering wheel detection problems into hand detection to determine whether driver's hands are on a steering wheel or not (the 1^{st} column) or if a cell-phone is being used (the 2^{nd} column).

eating, etc [10]. Everyday in the United States (U.S.), over 8 people are killed and 1,161 injured in crashes that are reported to involve a distracted driver. According to [6], there were 2,910 fatal crashes occurred on U.S. roadways that involved 2,959 distracted drivers, as some crashes involved more than one distracted driver, in 2013. Distraction caused by using a cell-phone while driving, i.e. 411 fatal crashes with 445 people died reported, is the most known example which significantly hinders driver awareness and reaction capabilities. Other secondary known examples of distraction are activities such as sending text messages, reading a book, eating, drinking, etc. In most cases of distractions, a driver just keeps only one hand or even no hand on the steering wheel. Therefore, successfully detecting a driver's

*These two authors contributed equally.

hands on the wheel aims at several goals. Firstly, in the interest of driver's safety, it provides the levels of attention of a driver to the road, i.e. using hands while operating the vehicle [15]. Secondly, it helps to analyze and understand driver behaviors, like maneuvering on a freeway or turning in an intersection [7].

The Federal Highway Administration (FHWA) [3] is leading efforts to develop and deploy computer vision and machine learning based driver monitoring algorithms that could potentially be deployed in real-world scenarios for driver monitoring as part of a law enforcement effort. These algorithms could be also used to automatically annotate the videos that have been collected. FHWA recently commissioned an exploration project that challenges researchers in both universities and industries to develop an useful indicator of the driver's state to extract information regarding the driver's disposition, passenger information, and driving performance (including detecting drowsiness, cell-phone usage, head pose tracking, monitoring if the driver has both hands on the steering wheel, etc).

In this paper, we propose a Convolutional Neural Network (CNN) based approach to handle the problems of cell-phone usage detection and hands on wheel detection as shown in Fig. 1. Our proposed CNN-based method named Multiple Scale Faster-RCNN (MS-FRCNN) first detects and extracts the regions of interest (RoI), namely, hands, faces and steering wheels. Each RoI is assigned to one confidence score. In order to investigate if the cell-phone is being used or determinate if both hands are on steering wheel, the high confident scores together with following observations are used. Firstly, it always has only one steering wheel per frame as shown in VIVA database. Therefore, the RoI with the highest score is chosen in the steering wheel detection module. Secondly, it always has only one face per frame as shown in SHRP-2 database. Therefore, the RoI with the highest score is chosen in the face detection module. Finally, there are more than one hand per frame as in VIVA database, thus, RoIs whose scores higher than a threshold T are chosen in the hand detection module. The geometric information is then used to make a decision if the hand is on the steering wheel or not, i.e. investigating intersections between the RoIs of the hands and the RoI of the steering wheel. A similar method is also employed to check if the cell-phone is being used, i.e. the RoI of hand holding the cell phone is either on the left or on the right of the RoI of the driver's face.

This paper presents the following contributions :

- Propose Multiple Scale Faster-RCNN (MS-FRCNN) approach, an improvement of F-RCNN [18], to robustly detect small objects like hands and faces collected under various scales, poses and environmental conditions. The experiments show that our MS-FRCNN archives better performance than F-RCNN in both hands on wheel detection and cell-phone usage detection while remaining at similar testing cost.
- The defined framework also has the capability to robustly detect the steering-wheel, the faces and the hands in a unified model. Not only detecting hands in vehicle, but our proposed method is also able to determine if a driver is using a cell-phone or how many hands are keeping on the steering wheel.
- Compared against the state-of-the-art cell-phone usage detection[19], our proposed method achieves higher accuracy, is less time consuming and is landmarking-independent. Notably, facial landmarking is a very challenging problem and time consuming.

The rest of this paper is organized as follows. Section 2 reviews the prior research carried out on detecting the hands in vehicle, particularly focusing on hands on phone and hands on wheel. Section 3 describes the state-of-the-art Region-based Convolutional Neural Network (R-CNN) and its advanced algorithms. We also discuss drawbacks of the existed methods in scenarios of driver distraction detection. Section 4 details our proposed MS-FRCNN method and how to apply it solve the problems of cell-phone usage detection and hands on the wheel detection. Section 5 describes the databases, the experimental protocols and the results obtained by our proposed method. Section 6 presents our final conclusions.

2. Related Work

In this section, we review the previous work of driver monitoring and the specific problems of hands on steering wheel detection and cell-phone usage detection.

Regarding the hands on the steering wheel detection, the multimodal vision method [13] was presented to characterize driver activities based on head, eye and hand cues. The fused cues from these three inputs using hierarchical Support vector Machines (SVM) enrich the descriptions of the driver's state allowing for evaluation of driver performance captured in on-road settings. However, this method with a linear kernel SVM for detection focuses more on analyzing the activities of the driver correlated among these three cues. It does not emphasize the accuracy of hand detection of drivers in challenging conditions, e.g. shadow, low resolution, phone usage, etc. Ohn-Bar et al. [14] introduced a vision-based system that employs a combined RGB and depth descriptor in order to classify hand gestures. The method employs various modifications of HOG features with the combination of both RGB and depth images to achieve a high classification accuracy. However, in the context of this work, it is impossible to get both RGB and depth images in cars since these videos are usually recorded in low resolution under poor illumination. Mittal et al. [12]

presented a two-stage approach to detect hands in unconstrained images. Three complementary detectors are employed to propose hand bounding boxes. These proposal regions are then used as inputs to train a classifier to compute a final confidence score. In their method, the context-based and skin-based proposals with sliding window shape based detector are used to increase recall. However, these skin-based features cannot contribute in our presented problem since all videos are recorded under poor illumination and gray-scale level. Meanwhile, these proposed methods [22] [16], [21] for hand tracking and analysis are only applicable in depth images with high resolution. They are therefore unusable in the types of videos used in this work.

Regarding detecting cell-phone usage by a driver, there has been many approaches proposed including non-vision based and vision and machine learning based approaches. To estimate the distance of a cell phone in use from the car's center, Yang et al. [9] harnessed a cars stereo system and Bluetooth network in an acoustic based approach. Thus, their approach is able to determine the cell-phone in use is from a driver or not. Breed et al. [4] placed three directional antennas at various locations inside a car to monitor emissions from a cell-phone. To find the most likely location of a cell-phone being used, a correlation could be made. Zhang et al. [24] applied Hidden Conditional Random Fields (HCRF) model onto features extracted features from the face, mouth, and hand regions to determine if a driver is using cell-phone. In their approach, they used cascaded AdaBoost classifier with Haar-like features for face detection. They also used a simple color-based approach for mouth detection. For the detecting hand region, they incorporated both color and motion information. Artan et al. [25] adopted a series of computer vision and machine learning techniques for detection and classification. They first used a Deformable Part Model (DPM) to localize the windshield region. Then, they used a DPM based simultaneous face detection, pose estimation, and landmark localization algorithm to locate a region of interest around the face to check for the presence of a cell-phone. Finally, they employed a Support Vector Machine (SVM) to classify to determine if the driver was using a cell-phone or not.

Far different from those approaches, our proposed method uses only one unified deep learning-based model to robustly detect multiple types of objects to solve the problems of driver distraction detection and highway safety including steering wheel detection, face detection and hand detection. In our deep learning framework, the global and the local context features, i.e. multi scaling, are synchronized to Faster Regions Convolutional Neural Networks in order to robustly achieve semantic detection. Furthermore, we also incorporate feature maps from shallower convolution feature maps, i.e. conv3 and conv4, for ROI pooling in order to enhance the capability of the network which is able to detect lower level features. Furthermore, to avoid poor performance because "larger" features usually dominate the "smaller" ones, we apply L2 normalization to each tensor before concatenating ROI pooling tensor,

3. Background

Deep ConvNets [2] recently have significantly improved object detection and image classification accuracy. In this section, we review various well-known Deep ConvNets, namely, Region based Convolutional Neural Networks methods including R-CNNs [17], Fast R-CNNs [8], and Faster R-CNNs [18].

3.1. R-CNN

The Region-based Convolutional Neural Network [17] uses a deep ConvNet to recognize given object proposals. It achieves great accuracy but is time-consuming. It is first trained on object proposals and fine-tunes a ConvNet with softmax regression layer at last. Then by replacing the last layer with SVM and using the features from fine-tuned ConvNet, the system is further trained for object detection. Finally it performs bounding-box regression. The system takes a long time to extract features from each image and store the features in a hard disk, which also takes up a large amount of space. At test-time, the detection process takes 47s for one image (with VGG16, on a GPU) due to the slowness of feature extraction.

3.2. Fast R-CNN

The main reason for R-CNN being slow is that it processes each object proposal independently without sharing computation. Fast R-CNN [8] tries to share the features between proposals. At test-time, it only extracts features once per image and uses ROI-pooling to extract features from the convolution feature map for each object proposal. It also uses a multi-task loss, i.e. classification loss and bounding-box regression loss. Based on the two improvements, the framework is trained end-to-end. The processing time for each image significantly reduced to 0.3s.

3.3. Faster R-CNN

Fast R-CNN accelerates the detection network by the ROI-pooling layer. However the region proposal step is out of the network hence remains a bottleneck, resulting in sub-optimal solution and dependence on the external region proposal methods. Faster R-CNN [18] addresses this problem with the region proposal network (RPN). An RPN is implemented as a fully convolutional network to predict the object bounds and the objectness scores. It uses anchors with different scales and ratios to achieve translation invariance. The whole system can finish proposal and detection in 0.2 second using very deep VGG-16 model [20], since the RPN shares the full-image convolution features with the detection network.

3.4. Limitations of Faster R-CNN

Fast R-CNN [8] and Faster R-CNN [18] achieve state-of-the-art performance on PASCAL VOC datasets. They can detect objects such as persons, animals, or vehicles. These objects usually occupy the majority of an image. However, in our problem we are interested in detecting hands and faces, which are usually small and low resolution objects as shown in Fig. 8. The detection network in Faster R-CNN has trouble to detect such small objects as shown in the first row of Fig. 3 where Faster-RCNN cannot find the small hands. The reason is that the ROI-pooling layer builds features only from one single high level feature map. For example, the VGG-16 model does ROI-pooling from the 'conv5' layer, which has a overall stride of 16. When the object size is less than 16 pixels, the projected ROI-pooling region is less than 1 pixel in the 'conv5' layer even if the proposed region is correct. Thus the detector will have much difficulty to predict the object class and bounding box location based on information from only one pixel.

4. Our Proposed Approach

This section presents our proposed Multiple Scale Faster-RCNN (MS-FRCNN) approach to robustly detect challenging objects, i.e. steering-wheel, face and hand, in the videos and images collected in SHRP-2 [23] and VIVA Challenge [5] databases. Our approach aims at synchronizing both the global and the local context features to Faster RCNN to achieve semantic detection with the highest accuracy. The average feature for layers in Faster-RCNN are employed to augment features at each location.

The rest of this section will be presented as follows. Firstly, we overview our proposed MS-FRCNN approach in subsection 4.1. Then, subsection 4.2 presents how to synchronize multiple scale features. Subsection 4.3 details our implementation for the new normalization layer in the Caffe framework. Finally, subsection 4.4 presents our proposed MS-FRCNN approach to the problem of hands on wheel and hand on phone detection.

4.1. Multiple Scale Faster-RCNN (MS-FRCNN)

The sizes of hands and faces in the observed images and videos are usually low-resolution. Therefore, it is a challenging task for the standard Faster R-CNN to successfully detect these objects. The reason for this difficulty is that the receptive fields in the last convolution layer (conv5) in the standard Faster R-CNN is quite large. For example, given a hand ROI region of sizes of 64×64 pixels in an image, its output in conv5 only contains 4×4 pixels, which is insufficient to encode informative features.

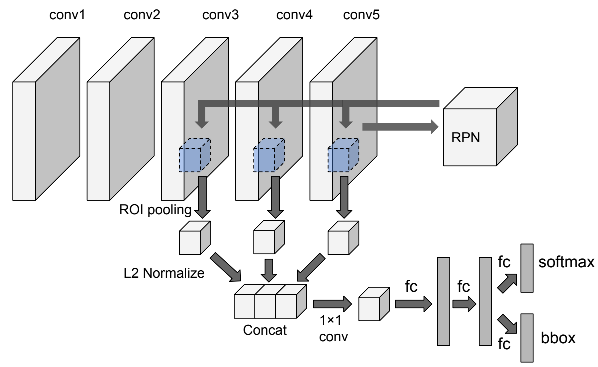

Figure 2. Our proposed Multiple Scale Faster-RCNN (MS-FRCNN) framework.

To make it even worse, as the convolution layers go deeper, each pixel in the corresponding feature map gather more and more convolutional information outside the ROI region. Thus, it contains higher proportion of information outside the ROI region if the ROI is really small. The two problems together, make the feature map of the last convolution layer less representative for small ROI regions.

Therefore, a combination of both global and local features, i.e. multi scaling, to enhance the global context and local information in the Faster RCNN network can help robustly detect our interested object. In order to enhance the capability of the network, we also incorporate feature maps from shallower convolution feature maps, i.e. conv3 and conv4, for ROI pooling (Fig. 2) So the network can detect lower level features which contain higher proportion of information in ROI regions.

In details, our approach keeps the same definition of the Regional Proposal Network (RPN) as in [18]. However, we define a more sophisticated network for Fast-RCNN to train these object proposals at various scales. Our defined network includes five sharing convolution layers, i.e. conv1, conv2, conv3, conv4 and conv5 as the standard one [18]. In the first two convolution layers, right after each convolution layer, there are one ReLU layer, one LRN layer and one Max-pooling layer respectively. In the next three convolution layers, right after each each convolution layer, there is only one ReLU layer. Especially, in three convolution layers, i.e. 3, 4 and 5, their outputs are also used as the input to three corresponding ROI pooling layers and normalization layers as shown in Fig. 2. These L-2 normalization outputs are concatenated and shrunk to use as the input for the next two fully connected layers. In the final steps, there are both a softmax layer for object classification and a regression function to take care of bounding box refinement.

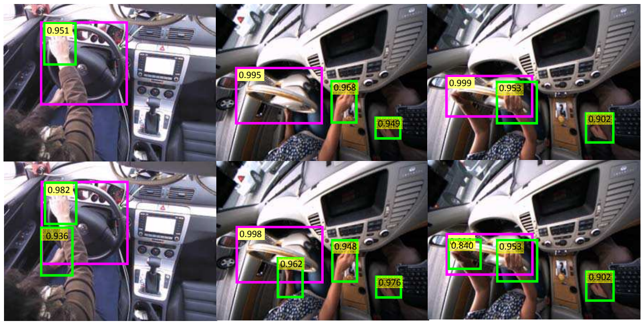

Figure 3. An comparison of our proposed MS-FRCNN against F-RCNN [18] on detecting small objects like hands. The 1^{st} row: F-RCNN [18]. The2^{nd} row: our MS-FRCNN

4.2. L2 Normalization

As discussed above and shown in Fig. 2, in order to expand the deep features of the defined objects at multiple scales, we need to combine three feature tensors after ROI pooling. In practice, the numbers of channels, scale of value and norm of feature map pixels are generally different at each layer, i.e the deeper layer usually has smaller-scaled values. Therefore, naively concatenating ROI pooling tensors usually leads to poor performance because the scale differences are too large for the weights in the following layers to readjust and tune. Then the "larger" features dominate the "smaller" ones and make the algorithm less robust.

Therefore, a straightforward solution for this problem is to normalize each ROI pooling tensor before concatenation [11]. Also, in this work, the system is able to learn the value of the scaling factor in each layer. This modification stabilizes the system and increases the accuracy.

Similar to the original work, we apply L2 normalization to each tensor. The normalization is done within each pixel in the pooled feature map tensor. After the normalization, scaling is applied on each tensor independently as:

$$
\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}
$$

$$
\|\mathbf{x}\|_2 = (\sum_{i=1}^{d} |x_i|)^{\frac{1}{2}}
$$

where the $\mathbf{x}$ and $\hat{\mathbf{x}}$ stand for the original pixel vector and the normalized pixel vector respectively. d stands for the number of channels in each ROI pooling tensor.

The scaling factor γ_i is then applied to each channel for every ROI pooling tensor:

$$
y_i = \gamma_i \hat{x}_i
$$

During training, the update for the scaling factor γ and input $\mathbf{x}$ is calculated with back-propagation and chain rule:

$$
\begin{aligned}
\frac{\partial l}{\partial \hat{\mathbf{x}}} &= \frac{\partial l}{\partial \mathbf{y}} \cdot \gamma \\
\frac{\partial l}{\partial \mathbf{x}} &= \frac{\partial l}{\partial \hat{\mathbf{x}}} \left(\frac{\mathbf{I}}{\|\mathbf{x}\|_2} - \frac{\mathbf{x}\mathbf{x}^T}{\|\mathbf{x}\|_2^3} \right) \\
\frac{\partial l}{\partial \gamma_i} &= \sum_{y_i} \frac{\partial l}{\partial y_i} \hat{x}_i
\end{aligned}
$$

Where $\mathbf{y} = [y_1, y_2, ..., y_d]^T$.

4.3. New Layer in Deep Learning Caffe Framework

In order to employ the L2 normalization, we need to integrate a Normalization layer into the current Faster-RCNN architecture. In our implementation, we follow the layer definition from ParseNet [11]. There are two more ROI pooling layers that extract features from the third and the forth convolution feature maps. The two ROI pooling layers, along with the original ROI pooling layer from the last, i.e. the fifth, convolution feature map, then pass the data through the normalization layer independently. The data were scaled to decent values and concatenated to a single tensor. We set the initial scaling factor to be 10 for all the three ROI pooling layers to ensure the downstream values in reasonable scales when training is initialized.

Next, the input to the fully-connected layer has to maintain the same sizes as the original architecture. Therefore, an additional 1×1 convolution layer is added into the network to compress the channel size of the concatenated tensor to the original one, i.e the same number as the channel size of the last convolution feature map (conv5), as shown in Fig. 2.

4.4. Our Proposed MS-FRCNN Approach to Driver Distraction

In this section, we describe how to use our proposed MS-FRCNN to solve the problems of driver distraction detection, i.e. cell-phone usage detection and hands on wheel detection. Unlike the previous hands on wheel detection approaches [13, 14, 12, 22, 16, 21], the proposed method first makes use of deep features extracted from our MS-FRCNN approach to individually detect the hand and the steering wheel. We then use geometric information, namely, the joint area between the detected hands and the detected steering wheel to decide how many hands are on the steering wheel. The flowchart of the hand on the wheel detection is given in Fig. 4.

In this method, we first apply the proposed MS-FRCNN approach to detect a steering wheel and hands separately. From the MS-FRCNN, we have a set of scores for the steering wheel detection and a set of scores for the hand detection. For the steering wheel detection, we choose a region

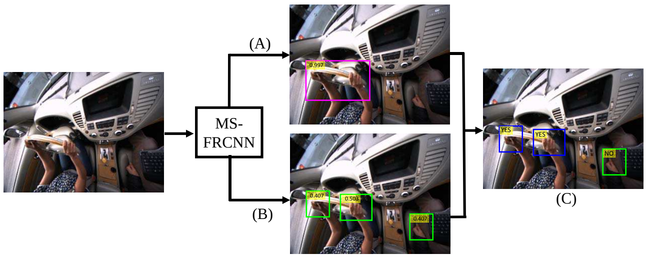

Figure 4. A flowchart of our hand on wheel detection by applying the proposed MS-FRCNN (A): Steering wheel detection, (B): Hand detection, (C): Hand on wheel detection

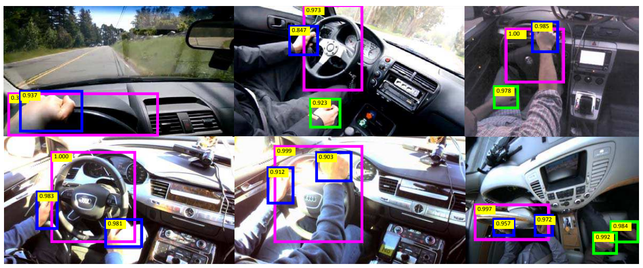

Figure 5. Some examples of our proposed hand on wheel detection method. The 1^{st} row: one hand is on the wheel. The 2^{nd}: both hands are on the wheel

that gives the best confidence score because of the following observation: there is always a steering wheel (whole or partial) on the videos from VIVA database. As for the hands detection, we choose regions whose probability scores higher than the threshold T, $T = 0.8$, according to our empirical results. The geometric information, namely, intersection between the detected steering wheel region and the detected hand regions is used to decide whether the hand is on the steering wheel or not. In our experimental results, if the joint region is bigger than 5% of the hand area then the hand is on the steering wheel. Some examples of our performance on the hands on the steering wheel detection including one hand (the first row) and both hands (the second row) on the steering wheel are given in Fig. 5.

Regarding cell-phone usage detection, we aim to detect the face and the hands instead of the cell-phone since the cell-phone is mostly occluded when being used. Furthermore, it cannot be seen even by human eyes because of poor illumination, low resolution. The flowchart of the cell-phone usage detection is given in Fig. 6. Similar to our previous experiment with the hands on the wheel detection, we apply our proposed MS-FRCNN approach to detect the face and the hands in an unified framework. As for the face detection, we choose a region with the best detection score. Indeed, although a driver's face maybe captured under different poses, it is always present in the videos (in SHRP-2 database) that we are interested in. Furthermore, we also observe that if the cell-phone is being used, the hand holding the cell-phone is always located on either left or right

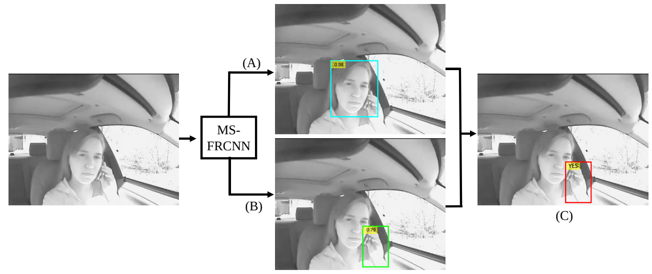

Figure 6. A flowchart of our cell-phone usage detection by applying the proposed MS-FRCNN (A): Face detection, (B) Hand detection, (C): cell-phone usage detection

Figure 7. Some examples of our cell-phone usage detection. The 1^{st} row: cell-phone is being used. The 2^{nd}: there is no cell-phone used

of the driver's face. Our decision if the cell-phone is used or not is made according to the geometric information (joint left or joint right) between the detected face and the detected hand. Some examples of our cell-phone usage detection is given in Fig. 7.

5. Experimental Results

5.1. Databases

The databases used in the experiments consist of a "drivers in the wild" database, i.e. Strategic Highway Research Program (SHRP-2) [23], collected by the Virginia Tech Transportation Institute (VTTI) [1], and hand database from Vision for Intelligent Vehicles and Applications (VIVA) Challenge [5]. By using these databases with numerous challenging factors, we aim to show the robustness and efficiency of our proposed method.

SHRP-2 Database: This database is collected by VTTI in order to evaluate the capability of safety driving system. In this collection, the platform was a 2001 Saab 9 - 3 equipped with two proprietary Data Acquisition Systems (DAS). These videos comprised of four channels of video, forward view, face view (resolution of 356×240), lap and hand view, and rearward view, recorded at 15 frames per second and compressed into a single quad video. These SHRP2 face view videos are used in our cell-phone usage detection experiments. We use the same training and testing datasets described in [19] which consists of 1,479 negative (no cell-phone) frames obtained 30 video segments of 11 subjects and 489 positive frames obtained from 20 video segments for training. The testing dataset consists of 9,288

Figure 8. Some examples of images in SHRP-2 database (the first row) and VIVA database (the second row).

negative frames and 3,735 positive frames of 30 subjects.

VIVA Hand Database: The dataset consists of 2D bounding boxes around hands of drivers and passengers from 54 videos collected in naturalistic driving settings of illumination variation, large hand movements, and common occlusion. There are 7 possible viewpoints, including first person view. In the challenging evaluation protocol, the standard evaluation set consists of 5,500 training and 5,500 testing images. To use this dataset for our hand on wheel detection experiment. Because there are two modules, namely, steering wheel detection and hand detection used to determine if a hand is on the wheel; however, the labeling data is only available for hand detection. Thus, we manually label steering wheels for both training and evaluating. We consider the accuracy of the proposed system under three cases: no hand is on the wheel, one hand is on the wheel and both hands are on the wheel. The testing data is categorized into three subsets containing 284 images without a hand on the steering wheel, 2,691 images with only one hand on the steering wheel and 2,525 images with both hands on the steering wheel. The groundtruth database will be publicly available. Some examples of images from SHRP-2 database and VIVA database are given in Fig. 8.

5.2. Experimental Results

In this section, we evaluate the effectiveness of the proposed MS-FRCNN method on two experiments: (1) cell-phone usage detection to answer the question that "Does a driver use a cell-phone while driving?" (2) hands on steering wheel detection to answer two questions that Q1: "Is there any hand on the steering wheel?" and Q2: "How many hands are on the steering wheel?". We evaluate the proposed method on the classification accuracy rates (Accuracy) and we consider the processing time by frame per second (FPS) metric. The Accuracy metric is computed based on the ratio between the number of samples correctly classified and the total number of samples. The proposed method is evaluated on a 64 bits Ubuntu 14.04 computer with CPU Intel(R) Core(TM) i7-4770K CPU@ 3.50GHz and Matlab 2014a.

Table 1 summaries the results obtained by the state-of-the-art [19], F-RCNN [18] and our approach MS-FRCNN on the cell-phone usage detection.

Table 1. Performance of [19], F-RCNN [18] and our approach on cell-phone usage detection with different metrics

Methods	FPS	Accuracy	Landmark
[19]	0.1	0.842	YES
F-RCNN [18]	0.04	0.924	NO
MS-FRCNN	0.06	0.942	NO

Table 2. Performance of F-RCNN [18] and our approach on the hand on the wheel with different metrics

Methods		Accuracy	FPS
F-RCNN[18]	Q1	0.91	0.04
	Q2	0.65	0.06
MS-FRCNN	Q1	0.93	0.09
	Q2	0.65	0.09

As we can see from Table 1, compared to the state-of-the-art method [19], our approach is not only independent to facial landmarking, but also achieves higher accuracy with less time consuming. Notably, facial landmarking is a very challenging problem. Following the same experiment setting up and running on the same training and testing, our MS-FRCNN gives better performance while having a similar processing time (Frame Per Second - FPS) compared against F-RCNN [18].

If low resolution and poor illumination are the big challenges in SHRP-2 database, under/over illumination and occlusion are the biggest obstacles in VIVA database. In our approach, the decision if the hand is on the steering wheel or not is made according to detected outcomes from hand detection and steering wheel detection together the geometric information between the detected hands and the detected steering wheel. We divide the hand on wheel detection into two sub-problems according to two above questions, i.e. Q1 and Q2. The answer for the first question Q1 is a binary classification (hand is presented on the wheel or not) whereas the answer for the second question Q2 is a 3-class classification (two hands/one hand/no hand). It is obviously that the second sub-problem is more challenging. The performance of our proposed MS-FRCNN compared against F-RCNN[18] on the hand on the steering wheel detection is given in Tables 2.

The cases when our hand on wheel detection algorithm fails can be divided into three categories: (1): occlusion (the 1^{st} row in Fig. 9): there is just very small portion of the hand is presented, thus, our method fails when detecting the hand; (2) overlapping (the 2^{nd} row in Fig. 9): Sometimes there are two hands on the wheel and they are overlapped, we just can see (detect) one hand, thus, our method fails when detecting the number of hands on the wheel ; (3) over/under illumination(the 3^{rd} row in Fig. 9): Under ugly environmental conditions where either steering wheel or hand cannot be seen, thus, our proposed algorithm fails to detect either hands or steering wheels.

Figure 9. The cases when our hands on wheel detection algorithm fails. 1^{st} row: hand occlusion; 2^{nd} overlapping; 3^{rd} under/over illumination

6. Conclusion

This paper presented an advanced deep learning based MS-FRCNN approach to effectively solve the problems of driver distraction monitoring and highway safety, namely, the hand on the wheel detection and the cell-phone usage detection. Our approach used the standard Region Proposal Network (RPN) generation incorporated feature maps from shallower convolution feature maps, i.e. conv3, conv4 and conv5 for the ROI pooling. The experiments conducted on VIVA and SHRP-2 databases showed our proposed approach obtained better accuracy, less testing cost and independent to facial landmarking compared to the state of the art [19], [18]. Additionally, our MS-FRCNN has archived higher accuracy while remaining at the similar cost comparing to Faster R-CNN.

References

[1] Virginia Tech Transportation Institute (VTTI). http://www.vtti.vt.edu/.

[2] I. S. A. Krizhevsky and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 201.

[3] F. H. Administration. https://www.fhwa.dot.gov/.

[4] D. S. Breed and W. E. Duvall. In-vehicle driver cell phone detector. http://www.google.com/patents/US8731530, May 2014. US Patent 8731530B1.

[5] N. Das, E. Ohn-Bar, and M. M. Trivedi. On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics. In *COnf. on ITS*, 2015.

[6] U. Department and T. N. H. T. S. Administration. Trafic safety facts - research note. http://www.distraction.gov/downloads/pdfs/Distracted_Driving_2013_Research_note.pdf.

[7] S. M. E. Ohn-Bar, A. Tawari and M. M. Trivedi. On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems. *Computer Vision and Image Understanding)*, pages 130–140, 2015.

[8] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[9] G. C. T. V. H. L. N. C. Y. C. M. G. J. Yang, S. Sidhom and R. P. Martin. Detecting driver phone use leveraging car speakers. In *MobiCom*, pages 97–108, 2011.

[10] A. Law Offices of Michael Pines. Distracted driving. https://seriousaccidents.com/legal-advice/top-causes-of-car-accidents/driver-distractions.

[11] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[12] A. Mittal, A. Zisserman, and P. H. S. Torr. Hand detection using multiple proposals. In *British Machine Vision Conf.*, pages 1–11, 2011.

[13] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi. Head, eye, and hand patterns for driver activity recognition. In *ICPR*, pages 660–665, 2014.

[14] E. Ohn-Bar and M. M. Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Transactions on ITS*, 15(6):2368–2377, 2014.

[15] T. H. Poll. Most u.s. driver's engage in ?distracting? behaviors: Poll. FMCSA-RRR-09-042, 2011.

[16] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *CVPR*, pages 1106–1113, 2015.

[17] T. D. J. M. R. Girshick, J. Donahue. Region-based convolutional networks for accurate object detection and semantic segmentation. *IEEE Transactions on PAMI*, Accepted may 2015.

[18] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[19] K. Seshadri, F. J. Xu, D. K. Pal, M. Savvides, and C. P. Thor. Driver cell phone usage detection on strategic highway research program (shrp2) face view videos. In *CVVT Workshop, CVPR*, 2015.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, pages 3213–3221, 2015.

[22] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *CVPR*, pages 824–832, 2015.

[23] E. The National Academies of Sciences and Medicine. The Second Strategic Highway Research Program (2006-2015) (SHRP-2). http://www.trb.org/StrategicHighwayResearchProgram2SHRP2/Blank2.aspx.

[24] F. X. Zhang, N. Zheng and Y. H. Visual recognition of driver hand-held cell phone use based on hidden crf. In *ICVES*, pages 248–251, 2011.

[25] R. P. L. Y. Artan, O. Bulan and P. Paul. Driver cell phone usage detection from hov/hot nir images. In *CVPRW*, pages 225–230, 2014.