Robust Hand Detection and Classification in Vehicles and in the Wild

T. Hoang Ngan Le KhaGia Quach Chenchen Zhu Chi Nhan Duong Khoa Luu Marios Savv
CyLab Biometrics Center, Carnegie Mellon University
{thihoanl, kquach, chenchez, chinhand, kluu, marioss }@andrew.cmu.edu

Abstract

Robust hand detection and classification is one of the most crucial pre-processing steps to support human computer interaction, driver behavior monitoring, virtual reality, etc. This problem, however, is very challenging due to numerous variations of hand images in real-world scenarios. This work presents a novel approach named Multiple Scale Region-based Fully Convolutional Networks (MS-RFCN) to robustly detect and classify human hand regions under various challenging conditions, e.g. occlusions, illumination, low-resolutions. In this approach, the whole image is passed through the proposed fully convolutional network to compute score maps. Those score maps with their position-sensitive properties can help to efficiently address a dilemma between translation-invariance in classification and detection. The method is evaluated on the challenging hand databases, i.e. the Vision for Intelligent Vehicles and Applications (VIVA) Challenge, Oxford hand dataset and compared against various recent hand detection methods. The experimental results show that our proposed MS-FRCN approach consistently achieves the state-of-the-art hand detection results, i.e. Average Precision (AP) / Average Recall (AR) of 95.1% / 94.5% at level 1 and 86.0% / 83.4% at level 2, on the VIVA challenge. In addition, the proposed method achieves the state-of-the-art results for left/right hand and driver/passenger classification tasks on the VIVA database with a significant improvement on AP/AR of 7% and 13% for both classification tasks, respectively. The hand detection performance of MS-RFCN reaches to 75.1% of AP and 77.8% of AR on Oxford database.

1. Introduction

The problems of hand detection and classification have been studied for years with the aim of ensuring the generalization of robust unconstrained hand detection algorithms to unseen images. However, the detection accuracy and classification score in recent hand detection systems [5, 3, 11, 10] are still far from achieving the same capabilities as a human due to a number of challenges in practice. For example, the hand variations, highly occlusions, low-resolution, strong lighting conditions, varied in shape and viewpoint as shown in Figure 1, are the important factors that need to be considered. Meanwhile, blurring of colors due to hand movement, skin tone variation in recorded videos due to camera quality are also the other difficulties in this problem.

(a) VIVA Hand database

(b) Oxford Hand database

Figure 1: Some examples of hand detection and classification (driver's left hand - RED, driver's right hand - GREEN, passenger's left hand - BLUE, passenger's right hand - YELLOW) results on VIVA database [2] (a); and hand detection Oxford database [14] (b) using our proposed MS-RFCN method (best view in color)

This paper presents an advanced Convolutional Neural Network (ConvNet) based approach, named Multiple Scale Region-based Fully Convolutional Networks (MS-RFCN), for hand detection and classification. In order to robustly deal with the challenging factors, we proposed to span the

receptive fields in the ConvNet in multiple deep feature maps. By this way, both global and local context information are able to be efficiently synchronized and simultaneously contribute to the human hand feature representation process. In particular, from the structure of R-FCN [4], we further introduce the Multiple Scale Regional Proposal Network (MS-RPN) to generate a set of region proposals and the Multiple Scale Region-based Fully Convolutional Neural Network (MS-FCN) to extract the regions of interest (RoI) of hand regions. Moreover, all trainable weight layers are convolutional and computed once for the entire image. Then, a bank of $k \times k$ position-sensitive score maps are generated for each category, and a $k^2(C+1)$-channel output layer with C object categories are computed. Finally, the class prediction is obtained by passing those score maps through an average pooling layer. The design of the proposed deep network can be seen in Figure 2. The deep learning Caffe framework [9] is employed in our implementation. The experiments are presented on the challenging hand databases, i.e. the Vision for Intelligent Vehicles and Applications (VIVA) Challenge [2] and Oxford Hand Detection database [14]. Our proposed method achieves the state-of-the-art results[1] in the problem of hand detection and classification on VIVA database.

The rest of this paper is organized as follows. In section 2, we review prior work on hand detection, the region-based CNN network in object detection and its limitations in the problem of hand detection and classification. Our proposed MS-RFCN approach for detecting and classifying hands from given input images in the wild will be detailed in section 4. Section 5 presents experimental results obtained using our proposed approach on the challenging hand database, i.e. VIVA challenge database and Oxford hand dataset. Finally, our conclusions on this work are presented in Section 6.

2. Related Work

Detecting, classifying and tracking of human hands have been widely addressed in many areas, such as: virtual reality, human computer interaction environment, driver behavior monitoring. In this paper, we focus on both problems of detecting and classifying hands in vehicles [2] and in the wild [14] detection. Indeed, a robust hand detection and classification system not only helps to study driver behavior and alertness but also provides document and human-machine interaction features, facilitates many tasks in human visual recognition, i.e. determining human layout and actions from still images. One of the first well performing approaches to detect the human hands was proposed by Mittal et al. [14]. They 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 a sliding window shape based detector are used to increase the recall. However, these skin-based features cannot contribute in our presented problem since all videos are recorded under poor illumination and gray-scale level. Later, Ohn-Bar et al. [16] 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. Ohn-Bar et al. [15] also introduced the multimodal vision method 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. Meanwhile, these proposed methods [23, 17, 22] 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. Unlike all the previous approaches that select a feature extractor beforehand and incorporate a linear classifier with the depth descriptor beside RGB channels, the state-of-the art hand detection approach named MS-Faster RCNN [11, 10] solves the problem under a deep learning framework where the global and the local context features, i.e. multi scaling, are synchronized to Faster Region-based Convolutional Neural Networks in order to robustly achieve semantic detection. In these methods, costly per-region subnetwork is applied hundreds of times and cannot take arbitrary input size.

To address these issues as well as provide the state-of-the-art performance on hand detection and classification with less time consuming, we proposed a Multiple scale Region-based Fully Convolutional Neural Networks (MS-RFCN). Our proposed MS-RFCN is fully convolutional with almost all computation shared on the entire image with position-sensitive score maps. Furthermore, the proposed method is able to adopt fully convolutional image classifier backbones, such as the latest Residual Networks (ResNet) for hand detection and classification.

3. Background

The recent studies in deep ConvNets have achieved significant results in object detection, classification and model-

[1] Submission date: March. 17rd, 2017, the VIVA hand detection ranking can be seen at http://cvrr.ucsd.edu/vivachallenge/index.php/hands/hand-detection/

ing. In this section, we review various well-known region-based Deep ConvNets. Then, we show the current limitations of the R-FCN, one of the state-of-the-art deep ConvNet methods in object detection, in the defined context of the hand detection and classification.

3.1. Region-based Convolutional Neural Networks

One of the most important approaches in the object detection task is the family of Region-based Convolutional Neural Networks. The first generation of this family, **R-CNN** [7], applies the high-capacity deep ConvNet to classify given bottom-up region proposals. Due to the lack of labeled training data, it adopts a strategy of supervised pre-training for an auxiliary task followed by domain-specific fine-tuning. Then the ConvNet is used as a feature extractor and the system is further trained for object detection with Support Vector Machines (SVM). Finally, it performs bounding-box regression. The method achieves high accuracy but is very time-consuming. The system takes a long time to generate region proposals, extract features from each image, and store these features in a hard disk, which also takes up a large amount of space. At testing time, the detection process takes 47s per one image using VGG-16 network [21] implemented in GPU due to the slowness of feature extraction.

R-CNN [7] is slow because it processes each object proposal independently without sharing computation. **Fast R-CNN** [6] solves this problem by sharing the features between proposals. The network is designed to only compute a feature map once per image in a fully convolutional style, and to use ROI-pooling to dynamically sample features from the feature map for each object proposal. The network also adopts 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.

Fast R-CNN accelerates the detection network using the ROI-pooling layer. However the region proposal step is designed out of the network hence still remains a bottleneck, which results in sub-optimal solution and dependence on the external region proposal methods. **Faster R-CNN** [19] addresses this problem by introducing the Region Proposal Network (RPN). A RPN is implemented in a fully convolutional style to predict the object bounding boxes and the objectness scores. In addition, the anchors are defined with different scales and ratios to achieve the translation invariance. The RPN shares the full-image convolution features with the detection network. Therefore the whole system is able to complete both proposal generation and detection computation within 0.2 seconds using very deep VGG-16 model [21]. With a smaller ZF model [24], it can reach the level of real-time processing.

Following similar object strategy in region-based systems [7, 6, 19], **R-FCN**[4] consists of two phases corresponding to region proposal and region classification. R-FCN consists of shared, fully convolutional architectures as the case of FCN [20]. However, it replaces the last few fully connected layers by convolutional layers to make efficient end-to-end learning and inference that can take arbitrary input size. To address a dilemma between translation-invariance in image classification and translation-variance in object detection, R-FCN uses a bank of specialized convolutional layers as the FCN output to construct position-sensitive score maps. Furthermore, R-FCN adopts the latest Residual Networks [8], for object detection by removing the average pooling layer and the fc layer and only use the convolutional layers to compute feature maps

3.2. Drawbacks of R-FCN in Hand Detection and Classification

The Region-based CNN family, e.g. R-CNN, Fast R-CNN Faster R-CNN [7, 6, 19] and the recent R-FCN [4] achieve the state-of-the-art performance results in object detection on the PASCAL VOC dataset. These methods can detect objects such as vehicles, animals, people, chairs, and etc. with very high accuracy. In general, the defined objects in this database often occupy the majority of an image, i.e. these objects have considerable numbers of pixels. However, when these methods are tested on the challenging Microsoft COCO dataset [12], the performance drops a lot, since images contain more small, occluded and incomplete objects. Similar situations happen in the problem of hand detection. We focus on detecting only hand regions that are sometimes small, low resolution as shown in Fig.1. The detection network in designed R-FCN is unable to robustly detect such tiny hands in car. The intuition point is that the Regions of Interest pooling layer, i.e. ROI-pooling layer, builds features only from the last single high level feature map.

Moreover, both RPN and R-FCN make predictions based on one single high-level convolutional feature map, i.e. conv4_23 whereas conv5_x layers are treated as fully connected layers in a convolutional style. The problem with conv4_23 is that it is a high-level feature with semantic information and its spatial resolution is 16 times smaller compared to the input image. Therefore, given a hand region with the sizes less than 16×16 pixels in an image, the projected ROI-pooling region for that location will be 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 the bounding box location based on information from only one pixel.

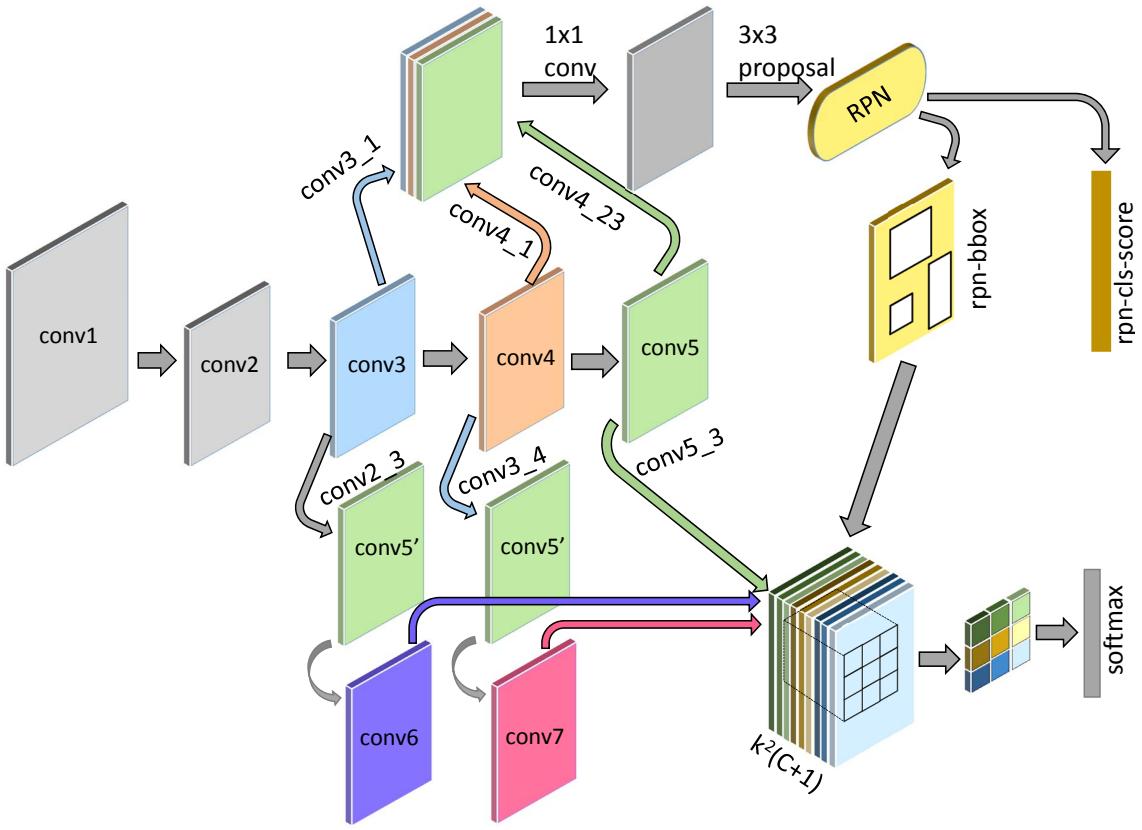


Figure 2: Our proposed Multiple Scale Region-based Fully Convolutional Networks (MS RFCN) approach to robust hand detection.

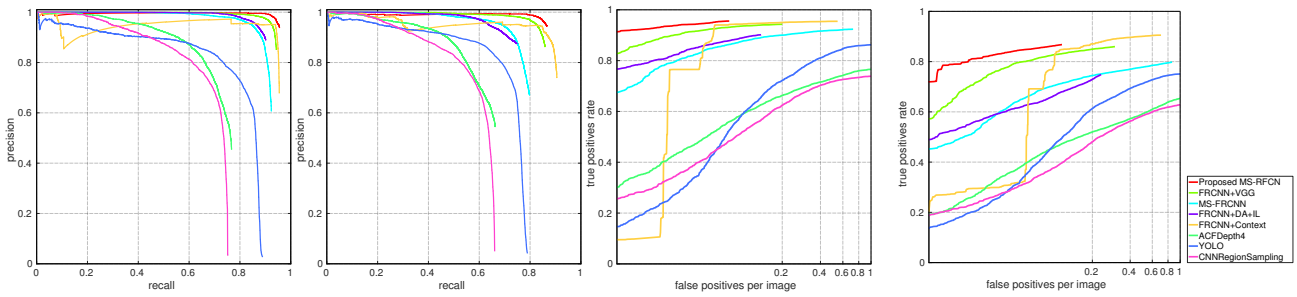
4. Our Approach to Robust Hand Detection and Classification

This section presents our proposed Multiple Scale R-FCN approach to robustly detect hand regions. Our approach utilizes the deep features encoded in both the global and the local representation for hand regions. Since the values of the filter responses range in different scales in each layer, i.e. the deeper a layer is, the smaller values of the filter responses are, there is a need for a further calibration process to synchronize the values received from multiple filter responses. The average feature for layers in R-FCN are employed to augment features at each location.

4.1. Deep Network Architecture

In problem of hand detection and classification, the sizes of human hand in observed images are usually collected under low-resolutions, strong lighting conditions varied in shape and viewpoint. It is a difficult task for the standard R-FCN to robustly detect these hand regions because the receptive fields in the last convolution layer in the standard R-FCN 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. When the convolution layers go deeper, each pixel in the corresponding feature map gather more convolutional information outside the ROI region. Thus, it contains higher proportion of information outside the ROI region if the ROI is very 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. multiple scales, to enhance the global and local information in the R-FCN model can help robustly detect and classify hand regions.

In order to enhance this capability of the network, we incorporate feature from multiple shallower convolution feature maps so that the proposed network inherits all the merits of both low-level localization information and high-level semantic information. Therefore, the network can robustly detect lower level hand features containing higher proportion of information in ROI region. Particularly, the defined network includes 101 convolution layers initialized using the pre-trained ResNets model [8]. Right after each convolution layer, there is a ReLU layer. But only 5 of these layers are followed with pooling layers that shrink the spa-



(a) L1-AP (b) L2-AP (c) L1-AR (d) L2-AR

Figure 3: ROC curves of hand detection on AP and AR measures obtained by FRCNN+VGG[1][1], MS-FRCNN [10], FRCNN+Context [1][2], FRCNN [25], ACF_Depth4 [5], [3], YOLO[18], CNNRegionSampling [3] and our proposed MS-FRCNN (solid, red) on VIVA database. (a): L1-AP, (b): L2-AP, (c): L1-AR, (d): L2-AR. Our method achieves the state-of-the-art hand detection results on this database.

tial scale. Therefore the convolution layers are divided into 5 major parts, i.e. conv1, conv2, conv3, conv4 and conv5. The first part, conv1, contains 1 convolutional layer. The second part, conv2, contains 3×3 convolutional layers. The third part, conv3, contains 4×3 convolutional layers. The fourth part, conv4, contains 23×3 convolutional layers. The fifth part, conv5, contains 3×3. And 1 more convolutional layer for average pooling. Specifically, for MS-RPN we consider the convolution layers conv3_1, conv4_1, conv4_23 with applying stride-2 pooling to conv3_1 to ensure them share the same spatial resolution. These convolution layers are then normalized (by L_2 norm) and concatenated together as the feature map for region proposal. For Ms-RFCN, we use the three convolutional layers, namely, conv2_3 and conv3_4 and conv5_3. In this network, we denote conv5'_x as a copy of conv5_x in the defined context that conv5_x and conv'5_x share the same architecture. The conv5'_x is then added after conv2_3 and conv3_4 to generate two new convolutional layers conv6_x and conv7_x, which are concatenated with the feature map from conv5_x.

Following the same strategy in region-based network, the proposed MS-RFCN consists of two phases corresponding to region proposal and region classification. Similar R-FCN, our network ends with a position-sensitive RoI pooling layer. This layer aggregates the outputs of the last convolutional layer and generates scores for each RoI which is classified into object categories and background. In MS-RFCN, all learnable weight layers are convolutional and are computed on the entire image. With end-to-end training, this RoI layer shepherds the last convolutional layer to learn specialized position-sensitive score maps. At the last convolutional layer, a bank of $k \times k$ position-sensitive score maps for each category is produced and thus a k2(C + 1)-channel output layer with C object categories is generated. The position-sensitive RoI layer conducts selective pooling, and each of the $k \times k$ bin aggregates responses from only one score map out of the bank of $k \times k$ score maps. The architecture of our proposed MS-RFCN for hand detection and classification is given in Fig. 2 where k is set as 3.

4.2. Deep Network Implementation

In our deep network architecture, features extracted from different convolution layers cannot be simply concatenated [13]. It is because the overall differences of the numbers of channels, scales of values and norms of feature map pixels among these layers. The detailed research shows that the deeper layers often contain smaller values than the shallower layers. Therefore, the larger values will dominate the smaller ones, making the system rely too much on shallower features rather than a combination of multiple scale features causing the system to no longer be robust.

In order to solve this problem, we introduce a normalization layer to the CNN architecture [13]. The system takes the multiple scale features and apply L_2 normalization along the channel axis of each feature map. Then, since the channel size is different among layers, the normalized feature map from each layer needed to be re-weighted, so that their values are at the same scale. After that, the feature maps are concatenated to one single feature map tensor. This modification helps to stabilize the system and increase the accuracy. Finally, the channel size of the concatenated feature map is shrunk to fit right in the original architecture for the downstream fully-connected layers.

Before normalization, all feature maps are synchronized to the same size so that the concatenation can be applied. In the MS-RPN, shallower feature maps are followed by pooling layers with certain stride to perform down-sampling. In the detection network, the ROI pooling layers already ensure that the pooled feature maps are at the same size. The implementation of L_2 normalization layer follows the layer definition in [13], i.e. the system updates the re-weighting factor for each feature map during training. In our archi-

tecture, we combine feature maps from three layers, i.e. conv3_1, conv4_1, conv4_23 in MS-RPN and conv5, conv6, conv7 in MS-RFCN, of the convolution layers. They are normalized independently, re-weighted and concatenated. The initial value for the re-weighting factor needs to be set carefully to make sure the downstream values are at reasonable scales when training is initialized. Additionally, in order to shrink the channel size of the concatenated feature map, a 1×1 convolution layer is then employed.

5. Experimental Results

This section first introduces the two experimental databases. Then, the evaluation protocols used in the experiment are described. Finally, the empirical results on hand detection and classification on each challenging database are detailed

5.1. Database Collection

VIVA hand database

The Vision for Intelligent Vehicles and Applications Challenge [2] 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. Some of the data was captured in test beds, while some was kindly provided by YouTube. In the challenging evaluation protocol, the standard evaluation set consists of 5,500 training and 5,500 testing images.

Oxford hand dataset The Oxford hand dataset [14] contains images collected from various different public image data set sources with no restriction was imposed on the pose or visibility of people, nor was any constraint imposed on the environment. In each image, all the hands that can be perceived clearly by humans are annotated. For this dataset, hand instances larger than a fixed area of bounding box (1500 sq. pixels) are considered 'big' enough for detections and are used for evaluation. This gives around 4,170 high quality hand instances. A training dataset is collected from six sources including Buffy Stickman, NRIA pedestrian, Poselet (H3D), Skin Dataset, PASCAL 2007, PASCAL 2012 with 9,163 instances and 2,863 big instances. The testing dataset contains 1,856 instances and 649 big instances from movie dataset, i.e. 'Four weddings and a funeral', 'Apollo 13', 'About a boy' and 'Forrest Gump'. The validation dataset contains 2,031 instance and 660 big instances from PASCAL 2007 and PASCAL 2012.

5.2. Evaluation Methods

To evaluate the performance on VIVA and Oxford databases, we compute the average precision (AP), average recall (AR) rate, and frame per section (PFS). AP is the area under the Precision-Recall curve whereas AR is calculated over 9 evenly sampled points in log space between 10^{-2} and 10^{0} false positives per image. A hand detection is considered true or false according to its overlap with the ground-truth bounding box. A box is positive if the overlap score is more than 0.5. The overlap score between two boxes is defined as $\frac{GT \cap DET}{GT \cup DET}$, where GT is the axis aligned bounding rectangle around area ground-truth bounding box and DET is the axis aligned rectangle around detected bounding box. The hand detection challenge is evaluated on two levels: Level-1 (L1): hand instances with minimum height of 70 pixels, only over the shoulder (back) camera view. Level-2 (L2): hand instances with minimum height of 25 pixels, all camera views. The left-right (L-R) and driver-passenger (D-P) hand classification are measured by AP and AR. 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: Performance of hand detection on AP, AR, FPS at both levels (L1 and L2) obtained by FRCNN+VGG[1][1], MS-FRCNN [10], FRCNN+Context [1][2], FRCNN [25], ACF_Depth4 [5], YOLO[18], CNNRegionSampling [3] and our proposed MS-FRCNN on VIVA database

Methods	L1-AP	L2-AP	L1-AR	L2-AR	FPS
[3]	66.8	57.8	48.1	36.6	0.783
[18]	76.4	46.0	69.5	39.1	35
[5]	70.1	60.1	53.8	40.4	–
[25]	90.7	55.9	**86.5**	53.3	–
[1][1]	89.5	86.0	73.4	64.4	4.9
[10]	92.8	82.8	84.7	66.5	0.234
[1][2]	93.9	91.5	85.2	77.8	6.32
Ours	**95.1**	**94.5**	86.0	**83.**4	4.65

5.3. Hand Detection and Classification on VIVA database

Table 1 summaries the performance of our proposed approach and the most recent approaches FRCNN+VGG[1], MS-FRCNN [10], FRCNN+Context [1], FRCNN [25], ACF_Depth4 [5], YOLO[18], CNNRegionSampling [3], and our proposed MS-RFCN Ap, AR and FPS at both levels (L1 and L2). Compare to the second best methods FRCNN+VGG[1], our proposed approach is higher 1.2% on L1-AP and higher than 0.8% on L2-AP whereas the AR obtained by our system is better from 3.0% to 5.6% on L1-AR, L2-AR, respectively. Processing time by our proposed method is near real with with 4.65 FPS on GPU. Fig. 3 visualizes the AP and the AR rates at both levels (L1 and L2). From Table 1 and Fig. 3, we can see that the proposed MS-RFCN outperforms others in higher AP, AR and less processing time.

Table 2 summaries the performance on AP, AR and

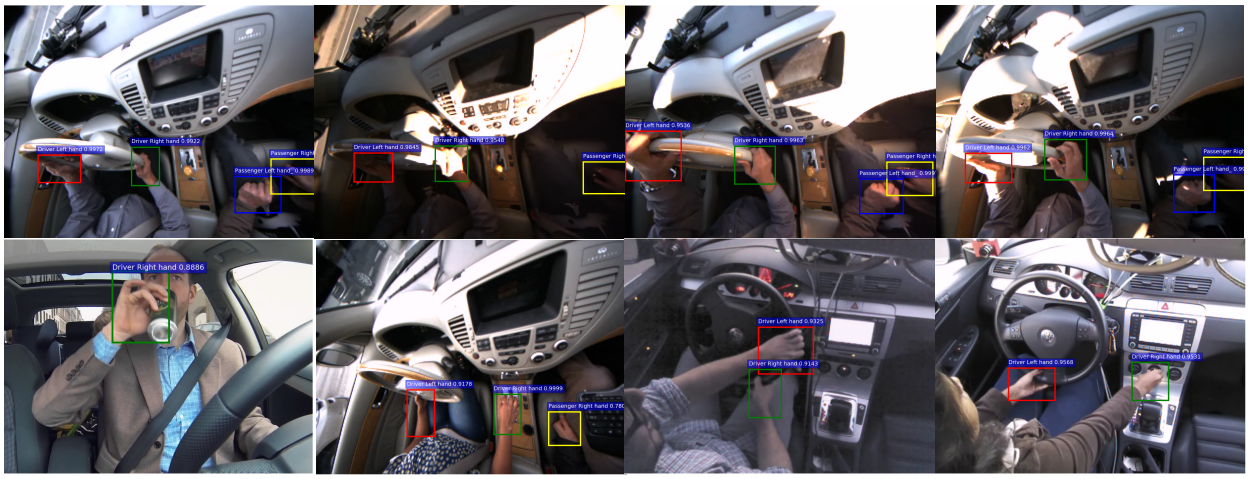


Figure 4: Some examples of hand detection and classification result using our proposed MS-FRCNN method on VIVA database [2]. The first row contains images from one video. The second row contains images from different videos

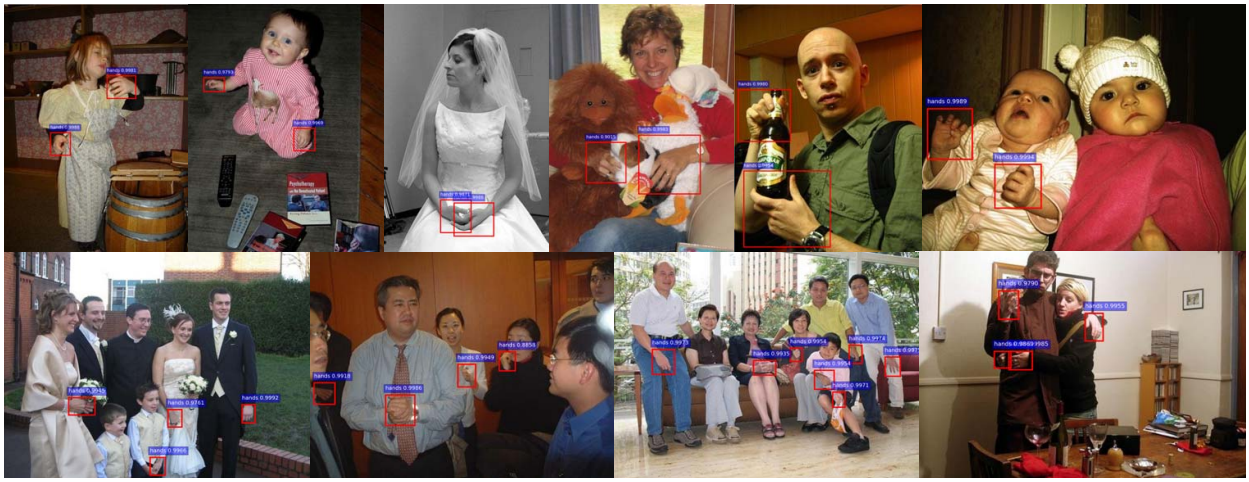


Figure 5: Some examples of hand detection result using our proposed MS-FRCNN method on Oxford database [14].

FPS of the most recent hand classification approaches FRCNN+VGG[1], ACF_Depth4 [5], CNNRegionSampling [3], and our proposed MS-RFCN. Compare to the second best proposed method FRCNN+VGG[1], our proposed approach achieves the state-of-the-art with higher 6.7% of AP and 6.8% of AR on the left/right hand classification and 13.2% of AP and 12.3% of AR on the driver/passenger hand classification on the VIVA database. Fig.4 visualizes some examples of hand detection and classification by our proposed MS-RFCN on VIVA database.

Table 2: Performance of hand classification (left/right(L/R) and driver/passenger (D/P) on AP, AR and FPS obtained by ACF_Depth4[5], CNNRegionSampling [3], FRCNN+VGG [1][1] and our proposed MR-RFCN on VIVA database

	L/R-AP	L/R-AR	D/P-AP	D/P-AR	FPS
[5]	47.5	33.7	43.1	30.5	11.6
[3]	52.7	42.3	57.3	47.3	0.783
[1][1]	68.6	63.0	57.7	53.3	4.9
Ours	**75.3**	**69.8**	**70.9**	**65.6**	4.65

5.4. Hand Detection on Oxford database

A. Mittal et al. [14] proposed different approaches for hand detection on Oxford database. Compare to the best performance using hand, context, skin score and post-processing obtained by [14], our performance achieves the state-of-the-art with 26.9% of AP and a comparable AR to [14]. Table 3 and Fig.5 visualizes some examples of hand detection by our proposed MS-RFCN on Oxford database.

Table 3: Performance of hand detection on AP and AR obtained by [14] and our proposed MR-RFCN on Oxford database

Methods	**AP**	**AR**
Mittal et al. [14]	48.2	85.3
Ours	**75.1**	77.8

6. Conclusion

This paper has presented our proposed MS-RFCN approach to robustly detect and classify human hand regions from images collected in vehicles and in the wild under various challenging conditions, e.g. highly occlusions, low resolutions, facial expressions, illumination variations, etc.The approach is benchmarked on two challenging hand databases, VIVA Challenge, and Oxford hand dataset, and compared against recent other face detection methods, e.g. FRCNN+VGG, MS-FRCNN, FRCNN+Context, etc. The experimental results show that our proposed approach consistently achieves the state-of-the-art results on both hand detection and hand classification .

References

[1] Hcii lab scut. 5, 6, 7

[2] *The Vision for Intelligent Vehicles and Applications (VIVA) Challenge, Laboratory for Intelligent and Safe Automobiles, UCSD*. http://cvrr.ucsd.edu/vivachallenge/. 1, 2, 6, 7

[3] S. Bambach, D. Crandall, and C.Yu. Viewpoint integration for hand-based recognition of social interactions from a first-person view. In *Proceedings of the 17th ACM International Conference on Multimodal Interaction (ICMI)*, pages 351–354, 2015. 1, 5, 6, 7

[4] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016. 2, 3

[5] N. Das, E. Ohn-Bar, and M. Trivedi. On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics. In *In IEEE Conf. Intelligent Transportation Systems*, pages 2953 – 2958, 2015. 1, 5, 6, 7

[6] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 3

[7] R. Girshick, J. Donahue, and J. M. T. Darrell. Region-based convolutional networks for accurate object detection and semantic segmentation. *IEEE Transactions on PAMI*, Accepted may 2015. 3

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 3, 4

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 2

[10] T. H. N. Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides. Robust hand detection in vehicles. In *ICPR 2016*, page accepted, Dec, 2016. 1, 2, 5, 6

[11] T. H. N. Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides. Multiple scale faster-rcnn approach to drivers cell-phone usage and hands on steering wheel detection. In *CVPRW 2016*, pages 46–53, June 2016. 1, 2

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. 2014. 3

[13] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 5

[14] A. Mittal, A. Zisserman, and P. H. S. Torr. Hand detection using multiple proposals. In *British Machine Vision Conference*, 2011. 1, 2, 6, 7, 8

[15] 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. 2

[16] 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. 2

[17] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *CVPR*, pages 1106–1113, 2015. 2

[18] S. G. R. Redmon, J.and Divvala and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*. 2016. 5, 6

[19] 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. 3

[20] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1605.06211, 2016. 3

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[22] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, pages 3213–3221, 2015. 2

[23] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *CVPR*, pages 824–832, 2015. 2

[24] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. 2014. 3

[25] T. Zhou, V. Pillai, P.J.and Yalla, and K. Oguchi. Hierarchical context-aware hand detection algorithm for naturalistic driving. In *ITSC*. 2016. 5, 6