Integrating Multiscale Deformable Part Models and Convolutional Networks for Pedestrian Detection

Wen-Hui Chen[a], Chi-Wei Kuan[b] and Chuan-Cho Chiang[c]
Graduate Institute of Automation Technology, National Taipei University of Technology, Taipei, Taiwan

Keywords: Advanced Driver Assistance Systems, Convolutional Neural Networks, Deformable Part Models, Pedestrian Detection.

Abstract: Pedestrian detection has many real-world applications, such as advanced driver assistance systems, security surveillance, and traffic control, etc. One of the pedestrian detection challenges is the presence of occlusion. In this study, a jointly learned approach using multiscale deformable part models (DPM) and convolutional neural networks (CNN) is presented to improve the detection accuracy of partially occluded pedestrians. Deep convolutional networks provide a framework that allows hierarchical feature extraction. The DPM is used to characterize non-rigid objects on the histogram of oriented gradients (HoG) feature maps. Scores of the root and parts filters derived from the DPM are used as deformable information to help improve the detection performance. Experimental results show that the proposed jointly learned model can effectively reduce the miss rate of CNN-based object detection models tested on the Caltech pedestrian dataset.

1 INTRODUCTION

According to the World Health Organization (WHO) report, road traffic accidents are the leading cause of death and injury. It indicated that several hundred thousand people lost their lives on roads each year (World Health Organization, 2013). The growth of automobiles over the past decade has contributed to the rise of the accident rate.

With the pervasion of digital technology, advanced driver assistance systems (ADAS) have gained popularity in the automotive industry in recent years. Nowadays, many modern vehicles have equipped with some sort of ADAS functions to provide drivers with safer, better, and more comfortable driving experience. Pedestrian detection is one of the key ADAS functions to safety control and collision avoidance.

The main challenges of pedestrian detection include human body articulation, occlusion, the changes of illumination and angle of view, and varying in appearance as well as scales. For example, people look different when wearing different clothes or taking different poses. In addition, lighting variations can also influence the image pixel values of an object, leading to the challenge of the object detection tasks in computer vision (Stefan Schnürle et al, 2017; Ortalda et al., 2018).

In the early stage, most detection algorithms were developed on designing handcrafted features. Deformable part models (DPM) that use the histogram of oriented gradients were the state-of-the-art approach in this period. Although many approaches have demonstrated some promising results, they were usually not robust and lack the generalized discriminative capability and some were even computationally intensive.

In 2012, the rise of convolutional neural networks (CNN) ignited the progress of object detection and became the mainstream of pedestrian detection research. In general, deep learning-based approaches have better performance over traditional learning models using handcrafted features on object detection but they heavily rely on training data to achieve qualified performance. In addition, most deep learning-based detection models were pre-trained on public datasets that lack occluded pedestrians, leading

[a] https://orcid.org/0000-0003-2015-5256
[b] https://orcid.org/0000-0003-2269-8158
[c] https://orcid.org/0000-0002-2535-6121

Chen, W., Kuan, C. and Chiang, C.
Integrating Multiscale Deformable Part Models and Convolutional Networks for Pedestrian Detection.
DOI: 10.5220/0009459005150521
In *Proceedings of the 6th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2020)*, pages 515-521
ISBN: 978-989-758-419-0

to a poor detection rate on partially occluded pedestrians.

In this study, we integrate the DPM and CNN to handle the occluding problem in pedestrian detection. The DPM is to provide a description of flexible human body models that can help detect partially occluded pedestrians.

The remainder of this paper is organized as follows. A review of the related work on pedestrian detection is provided in Section 2. The proposed approach and an overview of DPM are described in Section 3. The experimental setup and results are presented in Section 4. Finally, conclusions are drawn in Section 5.

2 RELATED WORKS

The study of pedestrian detection has been an active research topic for many years due to its potential applications. In this section, the related works on vision-based pedestrian detection are discussed. The development of pedestrian detection algorithms is closely related to the evolution of object detection as it is an application of computer vision. Numerous approaches have been proposed in the past decades, but some issues still remain unsolved and need to be addressed.

In the early stage, traditional pedestrian detection approaches highly rely on domain knowledge to design sophisticated features. Viola and Jones (2001) introduced a detection algorithm with Haar-like features and an AdaBoost cascade framework. The Viola-Jones (VJ) algorithm was originally proposed for face detection. Shortly afterward, it has been further applied to other detection problems, including pedestrian detection (Viola et al., 2003). The VJ detection algorithm was considered the first approach that can reach real-time performance, but it has some limitations including sensitive to lighting conditions and ineffective to encode the variance of a pedestrian in posture and appearance.

In 2005, the histogram of oriented gradients (HoG) was proposed as the features for human detection (Dalal & Triggs, 2005). This approach first divides an image into blocks and further divided each block into cells. The concept of HoG is to convert a pixel-based representation into a gradient-based one by calculating the gradient of each cell and building the histogram for all the orientations in a cell with discrete orientation bins.

HoG can be considered as the object information compressed and encoded in the orientation histogram that is ready to be fed into a classifier, such as support vector machines. However, HoG is very sensitive to image orientation and unable to characterize information well in a smooth region (Cheon et al., 2011).

The human body is a non-ridge object. Felzenszwalb et al. (2008) proposed deformable part models (DPM) that take the object deformation into consideration by including the deformable cost according to the displacement of each object part relative to its root location. DPM is a learning-based object detection algorithm defined by a constrained part-based model. It won the PASCAL VOC challenges (Felzenszwalb et al., 2010) and is considered the state-of-the-art algorithm for object detection before deep learning becomes popular.

During 2010-2012, the research of pedestrian detection reached a period of the plateau without significant performance improvement (Zhao et al., 2019). The reasons for the stagnant include: (1) sliding-based bounding boxes generation is inefficient and inaccurate; (2) manually designed features are not robust.

Since 2012, the resurgence of neural networks with deep architecture and the growth of computing power from hardware acceleration have led deep learning to great success in various fields, especially in computer vision. Instead of using handcrafted features, one of the advantages of deep learning is its ability to learn high-level features automatically.

In 2014, a pioneer work by applying deep learning to object detection was proposed (Girshick et al., 2014). The authors cleverly introduced the regions with CNN features to boost the mean average precision in object detection accuracy. Since then, the development of object detection begins to grow (Rasmussen et al., 2017). Now, there are many CNN-based object detection approaches including Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015), SSD (Liu et al., 2016), and YOLO (Redmon et al., 2016; Redmon & Farhadi, 2018). Deep learning to pedestrian detection in this period also achieved promising results (Sermanet et al., 2013; Ouyang & Wang, 2013; Luo et al., 2014). In addition to CNN, recurrent neural networks (RNNs) were also successfully applied to pedestrian detection to improve the average precision (Zhang & Kim, 2018).

One of the main challenges of pedestrian detection comes from the occlusion of the human body and has yet to discover an effective approach to deal with this issue. Ouyang and Wang (2012) proposed a DPM based method to improve the detection accuracy in occlusion. They used the restricted Boltzmann machine (RBM) to build the leaning model. However, the stacked RBMs as multiple layers of architecture lack semantic meanings.

Although deep architectures outperform shallow models in many challenging tasks, there are some ideas in state-of-the-art shallow models that are still useful and can be used to further improve the results achieved by deep learning models. In this study, we include the combined response derived from the DPM as the deformable score maps to improve the detection accuracy of the CNN-based models.

3 THE PROPOSED APPROACH

In this section, we describe how to improve the performance of pedestrian detection by adopting the deformable score maps from DPM and applying them to a CNN-based detection model. The computation flow of the proposed approach is illustrated in Figure 1.

At first, we collect training images with pedestrians and compute their transformed responses as deformable score maps from the DPM inference procedure. In this study, we use the Caltech pedestrian dataset to conduct the experiments. Then, we use the derived deformable score maps as additional information and feed them into the deep learning model in the learning stage to have the model learn the deformable information of a pedestrian.

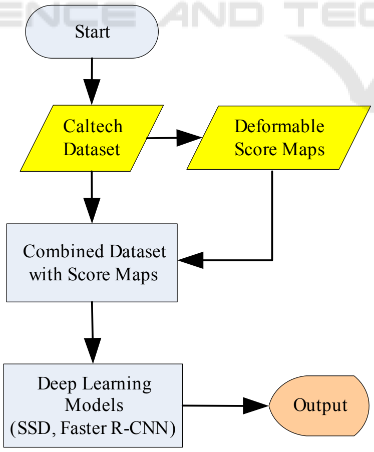

Start
Caltech Dataset
Deformable Score Maps
Combined Dataset with Score Maps
Deep Learning Models (SSD, Faster R-CNN)
Output

Figure 1: Flowchart of the proposed approach.

3.1 An Overview of DPM

The DPM recognizes objects involved with three major components: a root filter, a group of part filters, and a scoring mechanism. The root filter defines the detection window that covers the object to detect. Part filters are used to define a set of parts for the detected object. The connection between the root filter and its associated part filters is evaluated by a scoring mechanism that quantifies the spatial association with a deformable cost.

As the DPM learning model is built based on HoG features, the computation of HoG is the first step to the DPM inference. Here is the procedure to obtain HoG features of an image. Suppose that $l(x, y)$ is the pixel value at location (x, y). The gradient vector of the pixel (x, y) can be obtained as follows:

$$G_x(x,y) = l(x+1,y) - l(x-1,y) \tag{1}$$

$$G_y(x,y) = l(x,y+1) - l(x,y-1) \tag{2}$$

$$G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{3}$$

$$\theta(x,y) = \tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \tag{4}$$

where G_x and G_y are the partial derivatives on the x-direction and y-direction, respectively. The magnitude and direction of an image gradient vector are described in Equations (3) and (4).

Then, we divide the image into 8-by-8 pixel cells, where the magnitude values are stored cumulatively and added into nine bins for unsigned angles. Finally, we slide a block containing the size of 2-by-2 cells across the whole image. Histograms of the four cells in each block are concatenated into a vector and normalized with $L2$-norm. The HoG features are obtained by concatenating all the block vectors. As our target object is a pedestrian, we use eight part filters as components of a human body as illustrated in Figure 2.

To detect objects in various scales, we use image pyramids in four different scales as shown in Figure 3 to obtain a multi-scale DPM to represent the local shape descriptor generated by combining the histogram of edge orientations of each cell (Grauman & Darrell, 2005).

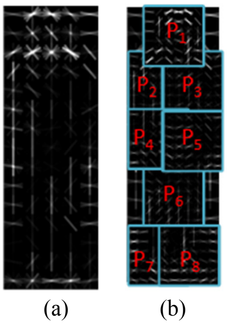


Figure 2: (a) The root filter and (b) part filters of an upright pose pedestrian in a DPM model.

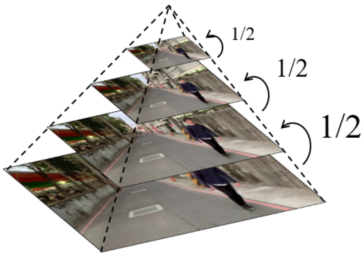


Figure 3: The pyramid of image. The full image resolution is taken at the bottom level.

3.2 Computation of Deformable Score Maps

The deformable scores in the DPM provide essential information for a possible displacement of object parts. It is beneficial to have this information added into a deep learning-based object detection models. The rest of this section describes the computation of deformable scores.

Suppose that the pedestrian model based on the DPM structure, M_p, is described by Equation (5).

$$M_p = (F_0, P_1, \ldots\ldots, P_N, B) \tag{5}$$

where F_0 is the root filter, N is the number of part filter, and B is the bias term. P_i is used to model part filters defined by a 5-tuple (F_i, v_i, s_i, a_i, b_i). The five elements of the tuple are listed below,

F_i: the i-th part filter

v_i: the box center of the i-th part relative to the root location

s_i: the box size

a_i, b_i: coefficients of a quadratic function measuring the deformable score of the i-th part

The response of a filter is computed by taking the dot product of the filter weights and the features in the HoG pyramid (Felzenszwalb et al., 2010; Cai, 2018). The deformable score is given by Equation (6).

$$\text{SCORE} = \sum_{i=0}^{N} F_i \cdot \phi(H, P_i) - \sum_{i=0}^{N} d_i \cdot \phi_d(dx_i, dy_i) + B \tag{6}$$

$$(dx_i, dy_i) = (x_i, y_i) - (2(x_0, y_0) + v_i) \tag{7}$$

$$d_i = (F_i, s_i, a_i, b_i) \tag{8}$$

$$\phi_d(dx_i, dy_i) = (dx_i, dy_i, dx_i^2, dy_i^2) \tag{9}$$

The first term $\phi(H, P_i)$ in Equation (6) describes a sub-window in the HoG pyramid H with the upper left corner in each part model P_i, while (dx_i, dy_i) in the second term describes the displacement of the i-th part and is defined in Equation (7). Hence, the deformable score can be expressed in terms of the dot product between d_i and ϕ_d defined in Equations (8) and (9), respectively. Figure 4 illustrates the procedure of deformable scores computation.

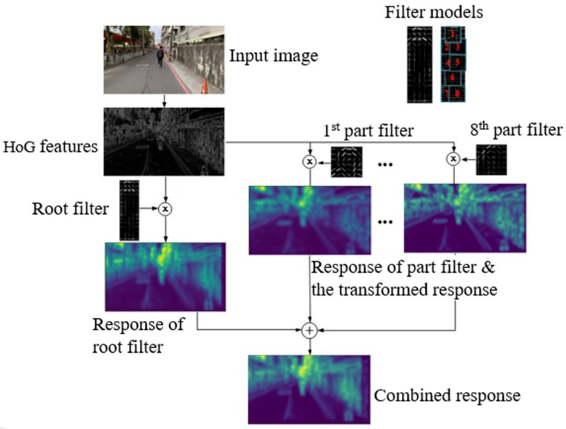


Figure 4: Illustration of computing deformable score maps.

The procedure of computing the deformable score maps is summarized as follows.

(1) Compute the pyramid of HoG features.

(2) Compute the response of the root filter by filtering the HoG pyramid features with the root filters leaned by the latent SVM.

(3) Compute the response of each part filter and the transformed responses.

(4) Sum up the response from multiple parts to obtain the combined response as deformable score maps

Figure 5 shows sample images from the Caltech dataset and their corresponding deformable score maps.

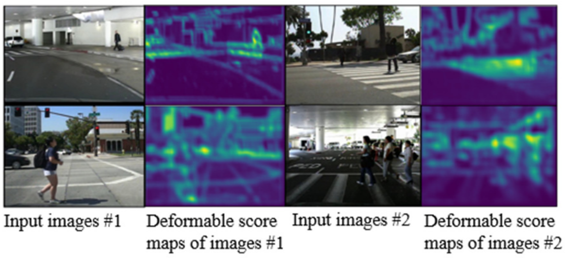

Input images #1
Deformable score maps of images #1
Input images #2
Deformable score maps of images #2

Figure 5: The first and the third columns show sample images from the Caltech dataset. The second and fourth columns show the corresponding score maps.

4 EXPERIMENTAL RESULTS

4.1 The Caltech Pedestrian Dataset

In this experiment, we use the Caltech pedestrian dataset (Dollar et al., 2012) to examine the proposed approach. The dataset contains videos approximately 10 hours collected from an urban traffic environment. There are about 2,300 pedestrians and a total of 350,000 bounding boxes labeled. The dataset was divided into 11 sessions, where six sessions for training and five sessions for the test.

The pedestrian instances are grouped into three scales: near (80 pixels or more in height), medium (30-80 pixels), and far (30 pixels or less). The subset *reasonable* in the dataset contains pedestrians that have 50 pixels or more in height. The training data in this study contains 128,419 images, while the test data contains 4,024 images.

In the training stage, we include the deformable score maps described in Section 3. Note that the original DPM filter parameters were trained on PASCAL VOC07. As the VOC07 dataset is lack of pedestrian samples, we re-trained DPM filter parameters with the Caltech pedestrian dataset.

4.2 Experimental Setup

We separate color channels of an image and obtain three corresponding score maps, namely *R*-score map, *G*-score map, and *B*-score map from *R*, *G*, and *B* channels, respectively. We conducted experiments and found that the inclusion of all the three score maps shows better results compared to the results derived only from a single channel score map or without score maps. Therefore, we include all the score maps derived from three channels in the training data.

CNN-based object detection models can be roughly divided into two categories: one-stage and two-stage frameworks. In order to verify the effectiveness of the proposed approach, two popular CNN-based detection models were used, namely SSD (one-stage model) and Faster R-CNN (two-stage model). The size of the root filter and part filters is set to 15×5 and 6×6, respectively throughout the experiments.

4.2.1 Tested on SSD

SSD is a popular one-stage detection model that has real-time detection ability without losing its detection accuracy. Unlike most detectors that run detection on the top layer of the network, SSD detects different object scales on different network layers. There are some variations in SSD. In this experiment, we use MobileNet-v2 as the lightweight backbone. The framework of detecting pedestrians is illustrated in Figure 6.

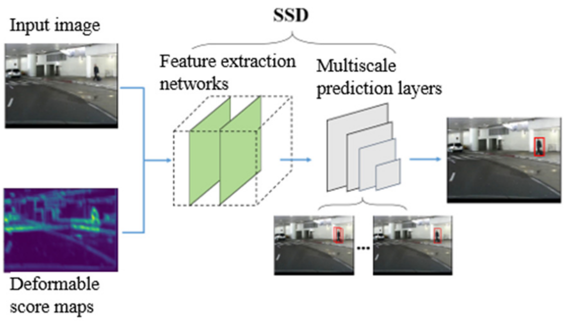


Figure 6: The framework of detecting pedestrians with deformable score maps on the SSD architecture.

When evaluating the proposed approach on SSD models, we observed that the performance in terms of miss rate is improved among all the pedestrian scales, as listed in Table 1. It has an improvement of 1.9% of miss rate in the scale of medium.

Table 1: Comparison of miss rate improvement on SSD models.

Scale / Models	Far	Medium	Near	Reasonable
SSD	96.0%	83.1%	29.4%	57.7%
Our Approach	95.8%	81.2%	28.8%	56.4%
Improvement	0.2%	1.9%	0.6%	1.3%

4.2.2 Tested on Faster R-CNN

Faster R-CNN is a popular two-stage detection model. We use ResNet50 as the base network for feature extraction. The framework of detecting pedestrians with deformable score maps on the Faster R-CNN architecture is illustrated in Figure 7.

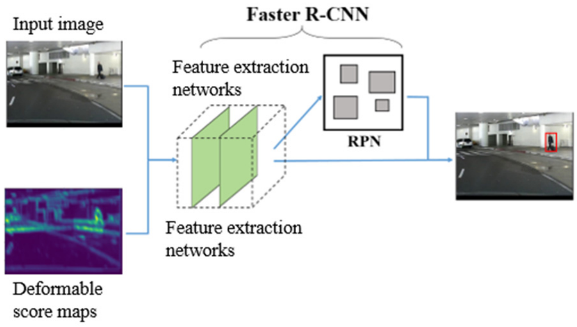


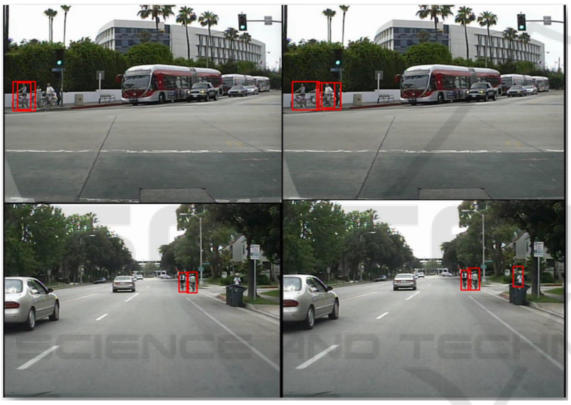
Figure 7: The framework of detecting pedestrians with deformable score maps on the Faster R-CNN architecture.

When evaluating the proposed approach on Faster R-CNN models, we observed that the miss rate is also improved among all the pedestrian scales, as listed in Table 2. It has an improvement of 2.7% of miss rate on the scale of near.

Table 2: Comparison of miss rate improvement on Faster R-CNN models.

Scale Models	Far	Medium	Near	Reasonable
Faster R-CNN	93.1%	57.8%	7.1%	25.3%
Our Approach	92.6%	56.8%	4.4%	22.8%
Improvement	1.5%	1.0%	2.7%	2.5%

Figure 8 shows examples of detection results. It can be seen that the proposed approach is able to detect partially occluded pedestrians that have been missed in the original Faster R-CNN model.



(a) (b)

Figure 8: Examples of detected results. (a) Detection results in Faster R-CNN models. (b) Detection results obtained by the proposed approach.

In order to evaluate the impact of occlusion percentage on detection accuracy, we further experiment with more occlusion setups and listed the results in Table 3 and Table 4 for SSD and Faster R-CNN, respectively.

From Table 3 and Table 4, it is clear that the proposed approach can help reduce the miss rate even in the case of heavy occlusion by providing the information of deformable scores. The miss rate was reduced from 1.3% to 2.6% in the Faster R-CNN architecture and from 1.7% to 2.1% in the SSD architecture. This is because non-rigid objects can be spatially organized in a deformable configuration of parts, and DPM is able to provide information that helps find partially occluded objects.

Table 3: Comparison of miss rate improvement at different percentages of occlusion on Faster R-CNN.

Occlusion Models	25%	50%	75%
Faster R-CNN	20.0%	24.9%	54.2%
Our Approach	17.4%	22.5%	52.9%
Improvement	2.6%	2.5%	1.3%

Table 4: Comparison of miss rate improvement at different percentages of occlusion on SSD.

Occlusion Models	25%	50%	75%
SSD	49.9%	57.7%	83.7%
Our Approach	47.8%	56.3%	82%
Improvement	2.1%	1.4%	1.7%

5 CONCLUSIONS

Pedestrian detection is an active research topic for the automotive and security industries. In this study, we have demonstrated a simple but effective jointly learned approach by including the deformable score maps derived from DPM into deep learning-based object detection models. The experimental results tested on the Caltech pedestrian dataset showed that the proposed approach is able to reduce a miss rate of 2.1% on SSD and 2.6% on Faster R-CNN in the case of 25% partial occlusion.

ACKNOWLEDGMENTS

This work is supported by the Taiwan Ministry of Science and Technology (MOST) under Grant No. MOST 108-2218-E-035-013

REFERENCES

Cai, Y., Liu, Z., Wang, H., Chen, X., & Chen, L. (2018). vehicle detection by fusing part model learning and semantic scene information for complex urban surveillance. *Sensors*, 18(3505). 1-19.

Cheon, M., Lee, W., Hyun, C. H., & Park, M. (2011). Rotation invariant histogram of oriented gradients. *International Journal of Fuzzy Logic and Intelligent Systems*, 11(4), 293-298.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern. Recognition.*

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4), 743-761.

Felzenszwalb, P., Girshick, R., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), 1627-1645.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. *IEEE Conference on Computer Vision and Pattern Recognition*.

Girshick, R. (2015). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition.*

Grauman, K., & Darrell, T. (2005). The pyramid match kernel: discriminative classification with sets of image features. *Proceedings of the Tenth IEEE International Conference on Computer Vision*.

Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. arXiv:1612.03144v2

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A. (2016). SSD: single shot multibox detector. *Proceedings of the 14th European Conference on Computer Vision.*

Luo, P., Tian, Y., Wang, X., & Tang, X. (2014). Switchable deep network for pedestrian detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ortalda, A., Moujahid, A., Hina, M., Soukane, A. & Ramdane-Cherif, A. (2018). Safe driving mechanism: detection, recognition and avoidance of road obstacles. *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*.

Ouyang, W., & Wang, X. (2012). A discriminative deep model for pedestrian detection with occlusion handling. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ouyang, W., & Wang, X. (2013). Joint deep learning for pedestrian detection. *Proceedings of the IEEE International Conference on Computer Vision*.

Rasmussen, C., Nasrollahi, K. & Moeslund, T. (2017). R-FCN object detection ensemble based on object resolution and image quality. *Proceedings of the 9th International Joint Conference on Computational Intelligence.*

Redmon, J., Divvala, S., Girshick, R., & Farhadi. A. (2016). You only look once: unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern. Recognition*.

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv:1804.02767v1

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*.

Schnürle, S., Pouly, M., vor der Brück, T., Navarini, A. & Koller, T. (2017). On using support vector machines for the detection and quantification of hand eczema. *Proceedings of the 9th International Conference on Agents and Artificial Intelligence*.

Sermanet, P., Kavukcuoglu, K., Pedestrian, S., & LeCun, Y. (2013). Pedestrian detection with unsupervised multi-stage feature learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern. Recognition*.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features, *Proceedings of the IEEE Conference on Computer Vision and Pattern. Recognition*.

Viola, P., Jones, M. J., & Snow, D. (2003). Detecting pedestrians using patterns of motion and appearance. *Proceedings of the Ninth IEEE International Conference on Computer Vision*.

World Health Organization. (2013). More than 270 000 pedestrians killed on roads each year. Retrieved Oct. 2019 from https://www.who.int/mediacentre/news/notes/2013/make_walking_safe_20130502/en/

Zhang, C. & Kim, J. (2018). Multi-scale spatial context features using 3-d recurrent neural networks for pedestrian detection. *Proceedings of the 21st International Conference on Intelligent Transportation Systems*.

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: a review. *IEEE Trans. On Neural Networks and Learning Systems*, 30(11), 3212-3232.