Journal of Computer Science

Original Research Paper

Enhancing Pedestrian Detection Using Context Information

Jorge Candido and Mauricio Marengoni

Universidade Presbiteriana Mackenzie, São Paulo, Brazil

Article history
Received: 3-04-2018
Revised: 12-06-2018
Accepted: 4-08-2018

Corresponding Author:
Jorge Candido
Universidade Presbiteriana Mackenzie, São Paulo, Brazil
Email:
Jorge.candido@mackenzie.com

Abstract: Detecting pedestrians among other objects in a digital image is a relevant task in the field of computer vision. This paper presents a method to improve the performance of a pedestrian detection algorithm using context information. A neural network is used to classify the region below pedestrian candidates as being floor or non-floor. We assume that a pedestrian must be standing on a floor area. This scene context information is used to eliminate some of the false-positive pedestrian candidates, therefore improving detector precision. The neural network uses 10 feature channels extracted from the original image to perform the region classification. This method may be used along with a large family of pedestrian-detecting algorithms. We used the ACF-LDCF algorithm to perform the tests in this research. The result shows that this method is very effective. We achieve a gain of 7% in ACF-LDCF algorithm performance on the Caltech pedestrian benchmark.

Keywords: Feature Extraction, Pedestrian Detection, Neural Network

Introduction

The pedestrian detection problem has attracted much attention in the field of computer vision due to its application in vision-aided Navigation Systems (INS) for automobiles (Panahandeh *et al*., 2012; Conrad and DeSouza, 2010), surveillance and elder people assistance. It is a challenging problem considering that some applications encompass great variations of illumination, pose and scale (Pears and Liang, 2001). Especially in outdoor scenes, those conditions are very unpredictable. The images in Fig. 1 show some of the challenges inherent to the pedestrian detection problem: great variation in scale, orientation, illumination and environment. The possibility of occlusion also brings some difficulties to the process.

For the last 15 years, pedestrian-detecting algorithms have made great strides. Some of these strides are due to the challenges posed by image databases used for training and testing (Dollár *et al*., 2009). The challenges imposed by the databases are catalysts for progress in some fields of computational vision.

Zhang *et al*. (2016) analyzes the main methods for pedestrian detection considered state of the art. The authors conclude that in spite of recent improvements, there is still room for progress and the use of information context is one of the most promising ways for achieving better results.

Context information is the data gathered from the image that do not directly belong to the searched object. Following Abowd *et al*. (1999), context information can be defined as "any information that can be used to characterize the situation of an entity, where an entity can be a person, place, or physical or computational object". The presence of a scene element can corroborate the algorithm response and help to gain confidence. This is the case of floor presence information applied to the pedestrian detection problem. With the exception of cases in which occlusion occurs, intuitively we can say that for each pedestrian on the scene there is a floor area just below.

In the work presented in Candido and Marengoni (2017), a neural network was used to classify an image area into floor or non-floor classes in order to perform a segmentation of the ground plane in outdoor images. Here the same idea was used to enhance pedestrian detector performance in an integrated system. The neural network analyzes the area below each pedestrian candidate selected by the detector algorithm and the response is combined with the candidate score and some of the false-positive candidates are eliminated in this process.

Extensive tests were applied along with the Caltech pedestrian database (Dollár *et al*., 2009). The Caltech-USA database is known as a benchmark in pedestrian detection algorithms and is widely used for training and testing. It consists of approximately 10 h of 30 fps video filmed from inside an automobile driving through urban regular traffic. Results show that the use of floor presence information can improve the accuracy of a pedestrian detection algorithm by up to 7%.

Fig. 1: Examples of pedestrians in outdoor scenes. Differences in pose, lighting and size pose great difficulties to the algorithm

The classifier was trained with a large number of example patches selected from images from the Caltech-USA database (Dollár *et al*., 2009). Positive examples were cropped from regions where pedestrians can walk: asphalt, sidewalk and grass. Negative examples are patches of buildings, cars, trees etc. The ANN was trained, tested and evaluated using those examples.

Related work

The majority of methods for pedestrian detection use one of the following approaches:

- Deformable part model
- Feature extraction and decision tree
- Convolutional Neural Network (CNN)

The following are some representative examples of works using each method.

The method based on DPM presented in Felzenszwalb *et al*. (2010), uses rigid root and deformable part filters but does not perform well on low- resolution objects. It was the basis for the development of the method called MT-DPM (Yan *et al*., 2013), that handles resolution differences between objects in the same image. The multi-resolution detection method is trained to learn structural commonness between samples of different resolutions.

Dalal and Triggs (2005), the authors show their studies with the use of locally normalized Histogram Oriented Gradients (HOG) in the task of pedestrian detection. HOG still remains as one of the most effective feature used in the classification task. Dollar *et al*. (2009) use the concept of feature channels based on an integral image (Viola and Jones, 2004) to gather the information that feeds a decision forest. Paisitkriangkrai *et al*. (2014), a method for feature extraction based on "spatial pooling" was proposed. The pooling operator selects a unique value that represents a region on the image. In that work, 2 types of image features were used: covariance matrix and Locally Binary Pattern (LBP).

One of the first examples of application with convolutional neural network was presented in Ouyang and Wang (2013). Despite the good results reached in this application, the structure was ineffective in dealing with variations in pedestrian scale. Du *et al*. (2017), the authors show a CNN structure called F-CNN. In this case, a first CNN generates a large number of candidates having a large number of false positives. That candidate list is applied to a series of CNNs working in parallel, each CNN specialized to detect pedestrian in a scale range.

The use of context information to aid pedestrian detection algorithms is not totally new. In the work presented in Jin *et al*. (2016), the authors use ground plane information to improve detector accuracy. Camera calibration parameters are used to constrict the floor geometry. Baek *et al*. (2016), a Bayesian learning process uses the location of the pedestrians on the training image group to define a Region of Interest (ROI) where the search is performed.

Proposed Method

Usually a pedestrian detector algorithm receives an image and performs an extensive search in order to locate each true-positive pedestrian that satisfies the search requirements (size, search area, occlusion level).

DOI: 10.3844/jcssp.2018.1074.1080

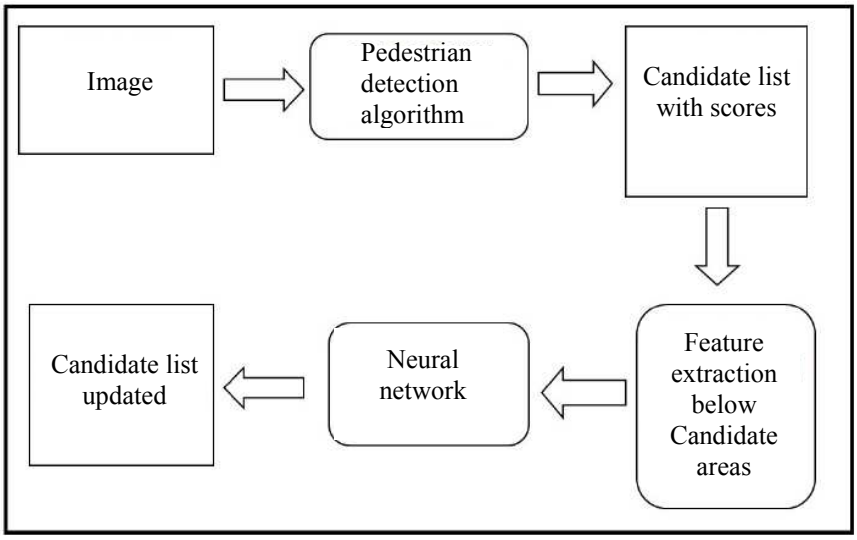

Fig. 2: Proposed system diagram

In that process, a large number of false-positive examples are included in the candidate list.

The higher the desired accuracy, the greater the number of false positives in the list, thus the lower the precision. The use of context information can improve this result.

In this work, we propose a novel system that integrates a pedestrian detector with a local context information extractor. Figure 2 shows the entire process described in a block diagram. The system applies the pedestrian detection algorithm to the image. The algorithm responds with a list of candidates in the form of localized bounding boxes and respective scores. At the next step, the system extracts the features below each bounding box and the neural network decides whether there is a floor area below each candidate. Candidates that are not standing on a floor area are eliminated from the list as a false-positive example.

Artificial Neural Network

The Artificial Neural Network used in this work is trained to perform a patch classification into 2 classes: floor and non-floor. The patch size is fixed at 16×16 pixels. This size was chosen through experiments with different patch sizes. Patches larger than 16×16 do not bring any improvement and demand much longer training time. However, patches smaller than 16×16 significantly degrade the result. The ANN is a feed forward-type network.

A typical application of ANN on computer vision uses pixel intensity values in the raw image feeding the ANN input (Rowley *et al*., 1998), but for texture classification, pixel values show a high level of redundancy. To avoid that effect, image features from the patches are used rather than the raw pixel value.

The activation function used for neurons in the hidden layer was the sigmoid. The output layer has only one neuron with a linear-type activation function which produces an output between 0 and 1. Therefore, responses below 0.5 represent the non-floor class, while responses equal or above 0.5 represent the floor class.

Feature Extraction

In the proposed system, the ANN input layer receives data from an intermediate step that transforms the original patch into a set of feature channels. Each channel represents the original patch in terms of a particular feature.

The calculation of the feature channels used in this work was inspired by Dollár *et al*. (2010). In that system, feature channels are used for pedestrian detection. We used the same feature combination resulting in 10 feature channels extracted from each patch. The feature channels are normalized gradient magnitude (1 channel), histogram of oriented gradient (6 channels) and LUV color channels (3 channels).

In the histogram of oriented gradient (HOG), the bins represent the orientation angles proposed for this application. The bins accumulate magnitude-weighted votes for gradients at the respective orientation. In our case, we use 6 different orientations, thus resulting in 6 different maps. The normalized gradient magnitude channel represents the actual value of the gradient intensity.

The 3 LUV color channels complete our set of features. Compared to other color space definitions, the LUV space delivered the best results in our experiments.

The same channel combination was used in Candido and Marengoni (2017) performing floor segmentation. In that work was demonstrated that the use of those 10 channels reached the best performance.

DOI: 10.3844/jcssp.2018.1074.1080

A pooling process is applied for each channel with the purpose of lowering data dimensionality and improving robustness of the system. In the maximum polling process used here, the highest value inside the window is used to represent the window region. A window size of 4×4 was used here.

Floor Detection

The objective of this work is to improve the performance of pedestrian detection algorithms by using context information of presence of floor area below each pedestrian candidate found by the detector algorithm. Floor area presence can improve confidence of the system by eliminating some false- positive examples found.

At the pedestrian detector output, a list of found pedestrians is presented with annotation of size and position in the form of a Bounding Box (BB). For each BB in the candidate list, 3 patches are cropped from the area below the BB following the schema shown in Fig. 3. The 3 patches are classified by the ANN and the results are combined to obtain evidence of a ground plane. If at least one patch is classified as floor, the detection is included in the final list. If no patch is classified as floor, the detection is eliminated from the list.

The annotated pedestrians in the Caltech training database were used to define the relative position of the patches to perform the floor detection. The ANN was used to analyze the region below each bounding box representing the pedestrians. The schema of the patches positions shown in the Fig. 3 was defined based on the statistical data gathered from those tests.

Fig. 3: Schema for patch extraction

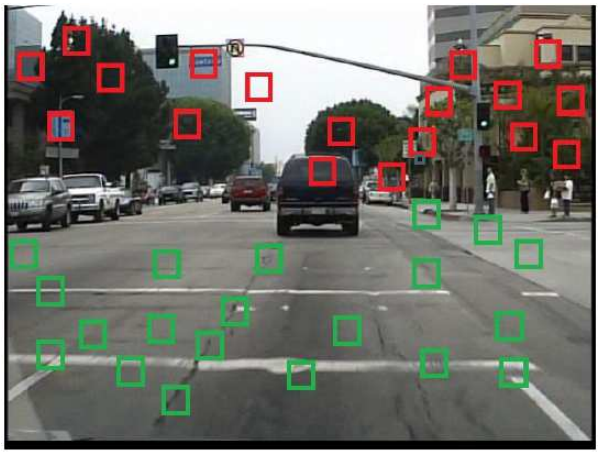

Fig. 4: Example of patch extraction. Red boxes are non-ground examples. Green boxes are ground plane examples

Experiments

During the experiments performed in this research, the influence of context information (floor presence) on the result of a pedestrian detection algorithm was evaluated. The approach consists of working with patches collected below the detected pedestrian to identify a floor area that reinforces confidence on the performed detection. The MATLAB Development Environment was used to test the algorithms and to carry out the image manipulations during all experiments performed.

Ground Plane Identifier Tool

The first step was create a ground plane identifier tool. This tool should work on classifying local patches of images between floor and non-floor classes. A neural network was chosen due to its simplicity and real time response. The training database was gathered from images of the Caltech pedestrian database. We cropped 1632 patches of floor examples and 4235 examples of non-floor examples. These patches were collected from 160 different images. Figure 4 shows an example of patches extracted from one image. The next step transforms each patch into a vector containing all extracted features. Finally, extra data is added to feature vectors defining the target label for use in the ANN training procedure. Value 0 is the label for non- ground examples and 1 is the label for ground examples.

A 5-fold cross-validation system was used in the neural network training following the schema of Fig. 5. The dataset was randomly divided into 5 subsets. In each trial, 4 folders are used for training and 1 is used for validation. By using this training method, the best neural network parameterization is more likely to be achieved, as per Haykin (1994). The training was executed by MATLAB's Resilient Backpropagation algorithm.

DOI: 10.3844/jcssp.2018.1074.1080

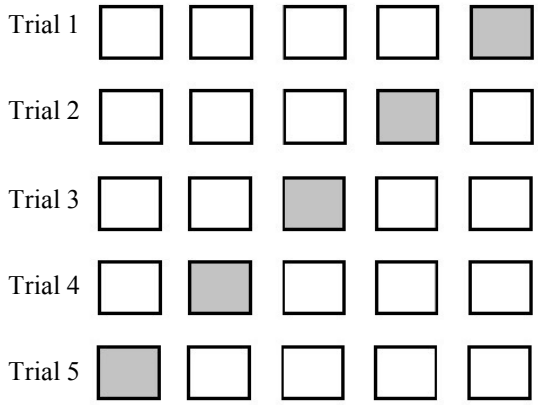

Fig. 5: 5-fold cross-validation training method. In each trial, four subsets are used for training and one subset (gray) is used for validation

The network structure that reached the best results has 15 neurons in the first hidden layer and 10 neurons in the second hidden layer. To prove the neural network efficiency, a second database was created having 113 positive patch examples (floor) and 117 negative patch examples (non-floor). Results of the neural network on this database give an idea of the efficiency of the tool as a ground plane identifier. The neural network classified correctly 100 floor patches (13 missed) and 112 non-floor patches (5 missed), reaching 92.17% accuracy.

However, these results are not very conclusive to evaluate the task of accurate extraction of ground presence below a detected pedestrian. This is mainly due to occlusion in the actual images. To get a more precise evaluation of this system, the following test was performed: for each pedestrian annotated on the images of the Caltech training group, the schema shown in Fig. 3 was applied. In this schema, the neural network analyzes 3 levels of patches below a Bounding Box delimiting a pedestrian. Floor presence is confirmed if at least one patch gives a positive answer. The result is that we could detect floor presence on 90.13% of pedestrians annotated on the Caltech training group.

Pedestrian Detection

The remainder of this section describes the tests for evaluating the whole system by integrating the pedestrian detector with context information (floor presence). Figure 2 shows a block diagram of the proposed system. The ACF-LDCF pedestrian detector algorithm (Nam *et al*., 2014) was used in the tests. The same process using the schema of Fig. 3, with 3 patches analyzed by the neural network, was used to detect the floor area below each candidate detected by the ACF-LDCF algorithm.

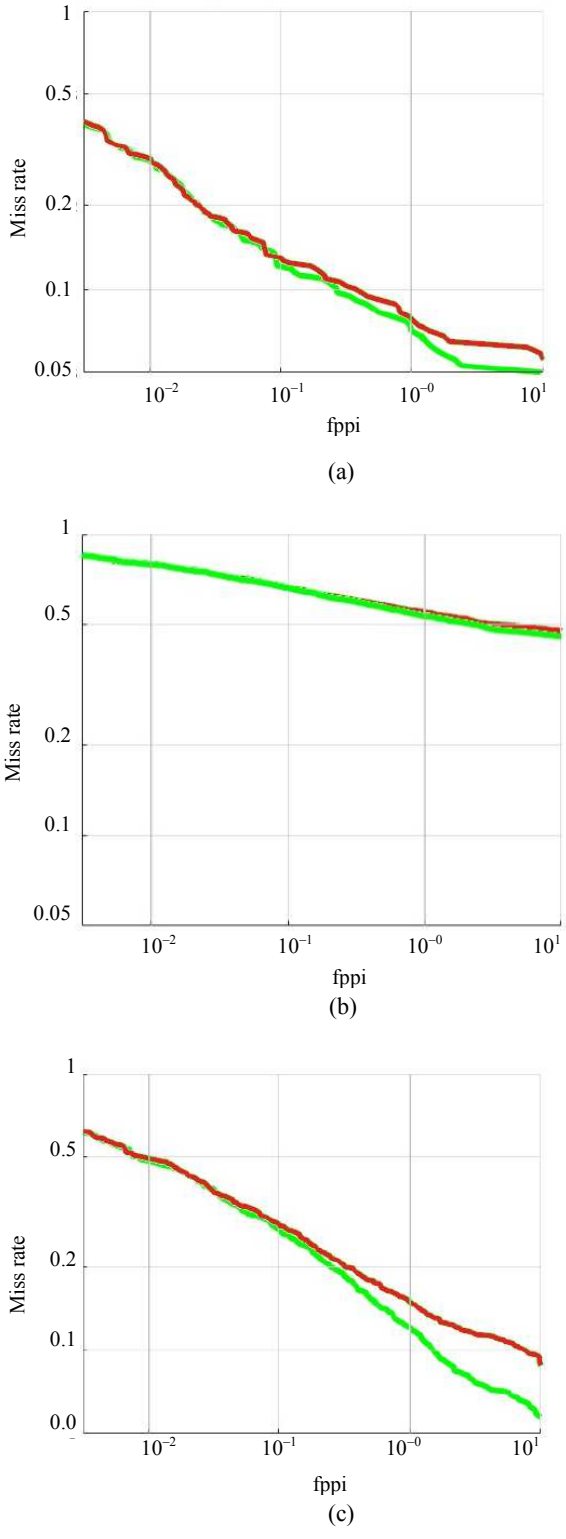

Fig. 6: ROC curves for the near sub-group (a), medium sub-group (b) and reasonable sub-group (c). The red trace shows the results for the original ACF-LDCF algorithm. The green trace shows the results for the proposed system

Table 1: Method performance comparison

Method	Recall	Precision
Jin *et al*. (2016)	-0.9%	+2.7%
Baek *et al*. (2016)	-2.8%	+19.5%
Ours	0	+ 7%

If at least one patch gives a positive answer for floor, the candidate is confirmed at the final detection list. If no patch gives a positive answer for floor, the candidate is possibly a false-positive and it is eliminated of the final detection list.

The system was used to detect pedestrians in images of 3 sub-groups of the Caltech dataset: the near sub-group, where pedestrians are 80 pixel or taller with no occlusion, the medium sub-group, where pedestrians are between 50 pixel and 80 pixel tall with no occlusion and the reasonable sub-group, where pedestrians are 50 pixel or taller with a low occlusion level.

The performance comparison between algorithms on the Caltech pedestrian benchmark is made using the ROC curve. The ROC curve shows the miss rate performance for the number of False Positives Per Image (FPPI). This comparison method is considered very effective because gives the number of object misses for each false-positive range. Figure 6 shows the comparison between ROC curves of the original ACF-LDCF algorithm (in red) and the proposed system (in green) on the 3 Caltech sub-groups aforementioned.

Another aspect of system performance relates to runtime. The use of context information caused a 55% increase in the algorithm's runtime. The original ACF-LDCF algorithm takes 1445 sec to run on the all 4000 images of the test group. Using context information along with the original algorithm, the runtime reaches 2239 seconds. Nevertheless, yet the system is capable of analyzing approximately 2 frames per second, making the system useful in many real-time applications.

We compare our method with 2 others that use context information to improve the performance of pedestrian detection algorithms. Jin *et al*. (2016), authors use the camera information to build a model and eliminate some false-positives based on this model. The method is capable of eliminating some false-positives reaching an improvement of 2.7% in the overall precision. But at the same time a few true- positives were eliminated. Consequently, the recall was degraded. Baek *et al*. (2016), authors use a Bayesian learning process to define a search area inside the image. This method eliminates 19.5% of the false-positives, but degrades the miss rate by 2.8%. The great advantage in this method is the computational time gain. Performing the search in a reduced area, this method requires only 70% of the computational time of a conventional method. Table 1 shows the performance comparison between these 2 method and ours.

Conclusion

In this study, we introduced a system integrating a pedestrian detection algorithm and an ANN for ground plane detection. The ANN helps the pedestrian detection algorithm by eliminating some false-positive examples on the pedestrian candidate list that do not have a visible ground plane below the bounding box that delimits the pedestrian. The ground plane identification works as context information that aids a pedestrian detection algorithm improving its performance.

In order to prove the effectiveness of the system, we perform tests in 3 sub-groups of the Caltech database. The best improvement (7.0%) was reached in the reasonable sub-group where pedestrians are 50 pixel or taller with a low occlusion level.

Results of this research indicate that the context information gathered from the scene aids the object detection task. The connection between the ground plane detected by the ANN and the actual pedestrians on the Caltech image database was proved by using the training image group where the location of pedestrians is known. The ANN was able to detect the ground plane for 90.13% of pedestrians in the images. In the future, this research will evaluate other elements in the scene such as cars, trees and other pedestrians, that could provide useful context information for the pedestrian detection task.

Acknowledgement

This paper was completed with the help and support of the Universidade Presbiteriana Mackenzie and especially acknowledgement of gratitude toward my teacher Mauricio Marengoni.

Author's Contributions

Jorge Cândido: Development of the main idea, experimental tests execution and text writing.

Mauricio Marengoni: General supervision and guidance.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and there are no ethical issues involved.

References

Abowd, G. D., A.K. Dey, P.J. Brown, N. Davies and M. Smith *et al*., 1999. Towards a better understanding of context and context-awareness. Proceedings of the International Symposium on Handheld and Ubiquitous Computing (HUC' 99). Springer, Berlin, Heidelberg, pp: 304-307.
DOI: 10.1007/3-540-48157-5_29

Baek, J., S. Hong, J. Kim and E. Kim, 2016. Bayesian learning of a search region for pedestrian detection. Multimedia Tools Appli., 75: 863-885. DOI: 10.1007/s11042-014-2329-z

Candido, J. and M. Marengoni, 2017. Ground plane segmentation using artificial neural network for pedestrian detection. Proceedings of the International Conference Image Analysis and Recognition, Springer, Cham, pp: 268-277. DOI: 10.1007/978-3-319-59876-5_30

Conrad, D. and G.N. DeSouza, 2010. Homography-based ground plane detection for mobile robot navigation using a modified EM algorithm. Proceedings of the IEEE International Conference on Robotics and Automation, May, 3-7, IEEE Xplore press, Anchorage, USA, pp: 910-915. DOI: 10.1109/ROBOT.2010.5509457

Dalal, N. and B. Triggs, 2005. Histograms of oriented gradients for human detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 20-25, IEEE Xplore press, USA, pp: 886-893. DOI: 10.1109/CVPR.2005.177

Dollár, P., C. Wojek, B. Schiele and P. Perona, 2009. Pedestrian detection: A benchmark. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 20-25, IEEE Xplore press, USA, pp: 304-311. DOI: 10.1109/CVPR.2009.5206631

Dollár, P., S. Belongie and P. Perona, 2010. The Fastest Pedestrian Detector in the West. In: Proceedings of the British Machine Vision Conference, Labrosse, F., R. Zwiggelaar, Y. Liu and B. Tiddeman, (Eds.). BMVA Press, pp: 68.1-68.11.

Dollar, P., Z. Tu, P. Perona and S. Belongie, 2009. Integral Channel Features. In: Proceedings of the British Machine Conference, A. Cavallaro, S. Prince and D. Alexander (Eds.), BMVA Press, pp: 91.1-91.11.

Du, X., M. El-Khamy, J. Lee and L. Davis, 2017. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. Proceedings of the IEEE Winter Conference on Applications of Computer, Mar. 24-31, USA, IEEE Xplore press, USA, pp: 953-961. DOI: 10.1109/WACV.2017.111

Felzenszwalb, P.F., R.B. Girshick, D. McAllester and D. Ramanan, 2010. Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Analysis Machine Intelligence, 32: 1627-1645. DOI: 10.1109/TPAMI.2009.167

Haykin, S., 1994. Neural Networks, a Comprehensive Foundation. Macmilan, ISBN-10: 0023527617

Jin, C., X. Cui, T. Woo and H. Kim, 2016. Method for pedestrian detection using ground plane constraint based on vision sensor. Proceedings of the International Conference on Electronics, Information and Communications, Jan. 27-30, IEEE Xplore press, Vietnam, pp: 1-4. DOI: 10.1109/ELINFOCOM.2016.7562937

Nam, W., P. Dollár and J.H. Han, 2014. Local decorrelation for improved pedestrian detection. Advances in Neural Information Processing Systems, pp: 424-432.

Ouyang, W. and X. Wang, 2013. Joint deep learning for pedestrian detection. Proceedings of the IEEE International Conference on Computer Vision, Dec. 1-8, IEEE Xplore press, Australia, pp: 2056-2063. DOI: 10.1109/ICCV.2013.257

Paisitkriangkrai, S., C. Shen and A. Van Den Hengel, 2014. Strengthening the Effectiveness of Pedestrian Detection with Spatially Pooled Features. In: Computer Vision – ECCV 2014, Fleet, D., T. Pajdla, B. Schiele and T. Tuytelaars (Eds.), Lecture Notes in Computer Science, Springer, Cham.

Panahandeh, G., N. Mohammadiha and M. Jansson, 2012. Ground plane feature detection in mobile vision-aided inertial navigation. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Oct. 7-12, IEEE Xplore press, Portugal, pp: 3607-3611. DOI: 10.1109/IROS.2012.6385503

Pears, N. and B. Liang, 2001. Ground plane segmentation for mobile robot visual navigation. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 29 Oct.-3 Nov., IEEE Xplore press, USA, pp: 1513-1518. DOI: 10.1109/IROS.2001.977194

Rowley, H.A., S. Baluja and T. Kanade, 1998. Rotation invariant neural network-based face detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 25-25, IEEE Xplore press, USA, pp: 38-44. DOI: 10.1109/CVPR.1998.698585

Viola, P. and M.J. Jones, 2004. Robust real-time face detection. International J. Computer Vision, 57: 137-154.

Yan, J., X. Zhang, Z. Lei, S. Liao and S.Z. Li, 2013. Robust multi-resolution pedestrian detection in traffic scenes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 23-28, IEEE Xplore press, USA, pp: 3033-3040. DOI: 10.1109/CVPR.2013.390

Zhang, S., R. Benenson, M. Omran, J. Hosang and B. Schiele, 2016. How far are we from solving pedestrian detection? Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 27-30, IEEE Xplore press, USA, pp: 1259-1267. DOI: 10.1109/CVPR.2016.141