Survey of pedestrian detection with occlusion

Chen Ning¹ · Li Menglu¹ · Yuan Hao¹ · Su Xueping¹ · Li Yunhong¹

Received: 7 March 2020 / Accepted: 22 September 2020 / Published online: 12 October 2020
© The Author(s) 2020

Abstract

Pedestrian detection is widely applied in surveillance, autonomous robotic navigation, and automotive safety. However, there are many occlusion problems in real life. This paper summarizes the research progress of pedestrian detection technology with occlusion. First, according to different occlusion, it can be divided into two categories: inter-class occlusion and intra-class occlusion. Second, it summarizes the traditional method and deep learning method to deal with occlusion. Furthermore, the main ideas and core problems of each method model are analyzed and discussed. Finally, the paper gives an outlook on the problems to be solved in the future development of pedestrian detection technology with occlusion.

Keywords Occlusion pedestrian detection · Neural network · Artificial features · Deep learning

Introduction

Pedestrian detection technology is a computer for the given video and image, to determine it is pedestrians, and mark the location of pedestrians. The rapid development of artificial intelligence technology also makes pedestrian detection set off a new upsurge in the field of computer vision. Pedestrian detection provides technical support and foundation for gait analysis, pedestrian identification, pedestrian analysis. These technologies are widely applied in video surveillance [1–4], self-driving cars [5–8], autonomous robots [9, 10] and many other fields.

The pedestrian detection technology has been advancing continuously in the past ten years. However, there is still a big problem to solve the occlusion situation. According to a recent survey, in a video that taken by a street, at least 70% [11] of pedestrians are occluded in Banks, shops, railway stations, and airports. The interference of complex background or other objects can increase the difficulty of pedestrian detection. At the same time, the commercial pedestrian detection system put forward high demands to overcome challenges.

Motivation

Pedestrian detection under occlusion has been widely used in the field of smart city. For example, vehicle-assisted driving systems, intelligent video surveillance, robotics, human—computer interaction systems, and security work all benefit from occluded pedestrian detection. In the field of intelligent transportation, assisted driving and autonomous driving are two important directions. Pedestrian detection under occlusion is one of the important foundations of the above directions. Accurate pedestrian detection under occlusion can help drivers to locate pedestrians and timely remind drivers to give way to people. At the same time, the detection results are helpful to risk management of driving behavior and improve driving safety. This has been playing an important role in ensuring the traffic safety of modern urban. In the field of security, it has become an important task to find the target under the occlusion by monitoring. Therefore, the research and summary of pedestrian detection under occlusion has far-reaching significance for both individuals and society.

In practical application, occlusion is common in crowded streets, railway stations and factories, and the pedestrian images under occlusion are in various shapes and forms. The accuracy of pedestrian detection algorithm will decrease when dealing with deformation and occlusion. The movement of pedestrians and the change of environment bring great challenges to the detection algorithm. Although the deep learning algorithms have made it great progress, it has

✉ Chen Ning
chennvictor@gmail.com

¹ School of Electronics and Information, Xi'an Polytechnic University, Xi'an, China

Fig. 1 Occlusion type

entered the bottleneck period due to the huge cost of training. Therefore, this paper first presents some previous successful cases, hoping to lay a foundation for future research on pedestrian detection under occlusion. Second, it summarizes and evaluates the current pedestrian detection algorithms under occlusion, hoping to bring some enlightenment to researchers and find new research hotspots for future research.

Previous work

Deformation and occlusion remain the main difficulties in pedestrian detection. Most previous studies focused on the advantages and disadvantages of pedestrian detection algorithms based on attitude deformation.

Pedestrian occlusion can be divided into two categories, one is the occlusion caused by background objects (inter-class), and the other is the occlusion caused by detection objects (intra-class), as it is shown in Fig. 1. The former kind is the difference between target and background, which often leads to the lack of target information. Furthermore, it leads to the missing of the object. The latter is the overlap between pedestrians, which often introduces a large amount of interference information. It leads to more virtual inspection. Pedestrian occlusion is divided into four levels according to the degree of occlusion between pedestrians [12]: 0, 1–35%, 35–80%, and above 80%. The research shows that the general pedestrian detection algorithm has good detection accuracy when the occlusion is between 0 and 10%.

Detection failure rate increases with the increase of the occlusion level. When the degree of occlusion exceeds 50%, pedestrians can hardly be detected.

The detection methods always followed the structure of “artificial feature + classifier” before the revolution of deep learning in computer vision. Deep Belief Network (DBN) [13], proposed by Geoffrey Hinton in 2006, is an extremely

Table 1 Different pedestrian detection categories with occlusion

Method	Traditional method	Deep learning
Algorithm framework	Artificial features (Haar, HOG, Edgelet, etc.) + classifier (SVM, Adaboost, etc.)	DBN \ RNN \ CNN \
Computational complexity	Low computational complexity	High computational complexity
Training sample demand	Fewer samples	More samples
Precision	Low	High

efficient learning algorithm. Since then, deep learning algorithms have blossomed in pedestrian detection.

Therefore, this paper divides the existing algorithms into two categories according to the detection framework: (1) Based on the traditional method [14, 15], and (2) Based on deep learning [16–18]. The traditional method includes hand-craft pedestrian features and classifiers, for example, Harr + Adaboost, Edgelet + Bayesian, HOG + SVM, etc. In traditional algorithms, there are two ways to deal with occlusion. One is based on a component detector. The other one is based on a special occluded classifier. The deep learning method relies on a neural network to learn pedestrian features autonomously. It has faster detection speed and higher detection accuracy; at the same time, it saves the time of manual feature selection. Table 1 shows the differences between the two categories. There are three mainframes of deep learning: (1) Deep Belief Network; (2) Convolutional Neural Network; and (3) Recurrent Neural Network. In deep learning algorithms, there are similar ideas to deal with occlusion. Some algorithms use the idea of the component detector due to their special structure of the neural network. Some algorithms use the optimization function to deal with occlusion. Figure 2 shows the key development of occluded pedestrian detection.

Traditional algorithm

Papageorgiou and Poggio proposed Haar in 2000. It can reflect the change of gray image scale, including four categories: edge feature, line feature, center-surround feature, and special diagonal line feature. Haar is the foundation of pedestrian detection technology.

The traditional detection methods always followed the structure of the “artificial feature + classifier”. First, the picture’s features should be extracted, including grayscale, edge, color, gradient histogram, and other information for the object. Then, the classifier determines which features belong to pedestrians. Such as, SVM, Adaboost, etc. The traditional method’s frame is shown in Fig. 3.

Fig. 2 Key development of occluded pedestrian detection **Fig. 3** Traditional method's frame

There are two main approaches to deal with occlusion in traditional detection methods: (1) The object is divided into different parts, and the visual part can infer the location of pedestrians. (2) A specific classifier is trained for the common occlusion in daily life to reduce the influence of occlusion and correctly judge the pedestrian position.

An algorithm based on the component detector

The component-based method is the most common and effective method to deal with the occlusion problem. The idea of this method is simple: though part of the pedestrian to be detected is occluded, the other parts can be used to locate the position of the pedestrian.

Leibe and Seemann [19] proposed a pedestrian detection algorithm in crowded scenes, which is equivalent to the prototype of pedestrian detection under occlusion. This kind of occlusion is an intra-class occlusion. The core part of their method is the combination of local and global cues via a probabilistic top-down segmentation. Mohan [20] found that if pedestrians are divided into four parts: head and shoulder, leg, left arm, and right arms, it is more effective to deal with occlusion. Mikołajczyk [21] further divided people into seven parts based on Mohan's method. Inspired by this, Bo and Nevatia [22] modeled humans as a collection of natural body parts, prompting Edgelet features. An Edgelet is a short segment of line or curve that denote the positions of normal vectors points in an Edgelet of $\{u_i\}_{i=1}^k$ and $\{n_i^E\}_{i=1}^k$, where k is the length of the Edgelet. Given an input image I , denote by $MI(\mathbf{P})$ and $NI(\mathbf{P})$ is the edge intensity and normal at position $\mathbf{p}$ of I . The affinity between the Edgelet and the image I at position w is calculated by the equator (1):

$$S(w) = (1/k) \sum_{i=1}^k M^I(u_i + w) \left| \left\langle n^I(u_i + w), n_i^E \right\rangle \right| \quad (1)$$

Xiaoyu takes HOG (Histograms of Oriented Gradients) and LBP (Local Binary Pattern) [23] as the feature set

and proposed a new human body detection method capable of handling local occlusion based on the component detector. Although part-based detectors perform better than other detectors, the sliding-window approach handles partial occlusions poorly. Two detectors are used to integrate the advantage of part-based detectors in occlusion handling to the sliding-window detectors: a global detector that scans the entire window and a partial detector in a local area. The response of the HOG-LBP feature of each block to the detector is used to construct an occlusion likelihood map. Once the occlusion is detected, part of the detector will be triggered to detect the visual part. Enzweiler and Eigenstetter [24] present a multi-cue component-based mixture-of-experts framework. Figure 4 shows the frame. The framework involves a set of component-based expert classifiers trained on features derived from intensity, depth and motion. This method, unlike Wu and Nevatia's approach. Wu requires specific camera settings, which need the camera to be positioned from top to bottom with the assumption that the heads of pedestrians in the scene were always visible of semantic segmentation. Flores-Calero [25] uses logic inference, HOG, and SVM are proposed to deal with occlusion. The input image is divided into twelve regions, and the feature vector is extracted for each region, and a classifier based on SVM has been built. These classifiers are used to build the final classifier. With this design, it is possible to capture the specific detail of each part of the human body, such as the head, legs, arms, and body.

Algorithm based on special occluded classifier

Training a set of special classifiers is another way to deal with occlusion. Each classifier is designed for a certain type of occlusion. Training special occluded classifier requires the prior knowledge of the occlusion types.

M. Isard found that adding the background appearance model into a pedestrian tracking algorithm is more robust

Fig. 4 Framework overview

and could effectively deal with deformation and occlusion. Wojek and Walk [26] apply the idea that not only individual pedestrians, but also surroundings need to be detected. They combined 3D scene tracking with detectors that perform occlusion handling by explicitly leveraging 3D scene information. The disadvantage of this approach, however, is that it is too costly. To solve this problem, Mathias and Benenson proposed Franken-classifiers [27]. It is less expensive to train a set of occlusion-specific classifiers. Sixteen occlusion-specific classifiers can be trained at only one-tenth of the cost of one full training. Felzenszwalb [28] proposed deformable part models (DPM). The algorithm adopts the improved HOG feature and uses SVM classifier and sliding-window detection, which is robust to the deformation of the target. Based on DPM (Deformable Parts Model), the model includes a linear filter incorporating a dense feature graph. A filter is a rectangular template defined by an array of d -dimensional weight vectors. The response, or score, of a filter F at a position (x, y) in a feature map G is the “dot product” of the filter and a subwindow of the feature map with a top-left corner at (x, y) :

$$\sum_{x', y'} F[x', y'] \cdot G[x + x', y + y'].$$

Andriluka and Schiele proposed a new two-person detector based on the DPM method to deal with occlusion. Instead of regarding the occlusion between people as interference, they think it is a peculiarity. This detector can predict the boundary boxes of two people with good results even under severe occlusion. The performance of this special occluded classifier is better than a single detector. However, the algorithm based on special classifier is time-consuming, and its robustness is not good. The algorithm does not work very well with a complex background.

Deep learning algorithm

There are three mainframes of pedestrian detection algorithms based on deep learning. (1) Based on depth belief network (DBN) [29]; (2) based on a convolutional neural network [30] (CNN); and (3) based on recurrent neural network (RNN). The Convolutional Neural Network is used widely

in the pedestrian detection algorithm. There are two ways to deal with occlusion in deep learning algorithm: One approach is to introduce the idea of part into a specific layer of the neural network; the other one is the optimization of neural network’s judgment mechanism.

Algorithm based on depth belief network

A deep belief network (DBN) proposed by Geoffrey Hinton in 2006 is an extremely efficient learning algorithm, which is a generic model. By training the weights among its neurons, we can let the whole neural network generate training data according to the maximum probability. In other words, pre-training + Fine-tuning. This idea has become the main framework of deep learning algorithms. The components of DBN are Restricted Boltzmann Machines (RBM). The process of training DBN is carried out layer by layer. In each layer, data vectors are used to infer the hidden layer, which is then treated as the data vector of the next layer (the higher layer).

Wanli and Xiaogang [31] combined the component model with DBN. They formulate feature extraction deformation handling, occlusion handling, and classification into a joint deep learning framework and propose a new deep network architecture. When part detection map and part scores are obtained, the joint framework can take full advantage of them. However, when there is an occlusion or large deformation, to integrate the fraction of partial detectors is a key problem to be solved urgently. In order to solve the defects of part detectors, they proposed a probability model based on improved RBM [32]. The hierarchical structure of the DBN model matches the multi-layers of the parts model well. This can achieve more reliable visibility estimation, and it is better to eliminate the influence of occlusion. The framework is shown in Fig. 5. It works well with both single-detector and multi-pedestrians systems.

Algorithm based on convolutional neural network

Convolutional Neural Networks (CNN) are a class of Feed-forward Neural Networks that contain convolutional com-

Fig. 5 Network framework

putation. Figure 6 shows the framework. Pedestrian detection algorithm based on Convolutional Neural Networks is mainly divided into two categories. First, it is the two-stage detector algorithm, which divides target recognition and target location into two parts. Region-Convolutional Neural Networks(R-CNN) series algorithm has high accuracy at a slower speed. Second, it is one-stage detector algorithm that includes Single-Shot MultiBox Detector (SSD) [33, 34] and You Only Look Once (YOLO) [35, 36]. YoLo is fast, but it has erratic effects with inherent advantages in detecting small targets and dense targets. SSD has high accuracy while maintaining fast speed.

At present literature, most of the pedestrian detection algorithms are based on a two-stage detector framework. Wanli [37] proposed deformable deep convolutional neural networks for generic object detection. The proposed algorithm has a new pre-training strategy to learn feature representations more suitable for the object detection task, which significantly improves the effectiveness of model averaging. Furthermore, jointly learning deep features [38], deformable parts, occlusion, and classification are proposed to established automatic mutual interaction among components. Yonglong Tian and Ping Luo [39] proposed the Deep-Parts, which is inspired by Franken-classifiers. Deep-Part introduces the idea of constructing a part pool that

covers all the scales of different body parts and automatically chooses important parts for occlusion handling. These methods' occlusion handling strategy is to learn a set of detectors and integrate the output of these ensemble models. But it is complicated and time-consuming. Shanshan Zhang combines Faster R-CNN with an attention mechanism [40]. This method is easy to train and has low overhead. The attention mechanism has been widely used in CNN for object detection. The additional attention mechanism guides the detector to pay more attention to visible body parts, as it is shown in Fig. 7.

Zou [41] proposed an attention guided neural network model (AGNN), which uses a fixed-size window slides on a still image without overlapping to generate a set of sub-images. The attention network performs local feature weighting by selecting the features of the pedestrian's body parts. Zhou and Yuan [42] propose a reduced computational complexity of a multi-label learning approach that jointly learn part detectors to capture partial occlusion patterns. The part detectors share a set of decision trees via boosting to exploit part correlations.

The introduction of part detectors to optimize loss function is a good strategy to deal with occlusion. Xinlong Wang and Tete Xiao [43] set repulsion loss function on the Faster R-CNN. This loss is driven by two motivations: the attraction by target, and the repulsion by other surrounding objects. The crowd occlusion makes the detector sensitive to the threshold of non-maximum suppression (NMS): a higher threshold brings in more false positives, while a lower threshold leads to more missed detections. The repulsion loss consists of two parts: the attraction term to narrow the gap between a proposal and its designated target, and the repulsion term to distance it from the surrounding non-target objects. Then, Shifeng and Longyin [44] propose a new aggregation loss function. The function enforces proposals to be close and locate compactly to the corresponding objects. At the same time, a new part occlusion-aware region of interest (PORoI) is proposed to replace the original RoI. PORoI can integrate the prior structure information of the human body with visibility prediction into the network to handle occlusion. Then, Cao, JL proposed location bootstrap and semantic transition, which is used to reweight regression loss and adds more contextual information and relieves semantic inconsistency of the skip-layer fusion. Sumi [45] proposed Frame-Level

Fig. 6 Convolutional neural network framework

Fig. 7 Flowchart of attention guided Faster R-CNN pedestrian detector

Fig. 8 LSTM decode

Difference (FLD) features, which will extract the features by finding the difference between the adjacent frame and retaining the noticeable differences. Using a combination of proposed features with other existing algorithms can improve the occluded pedestrian detection accuracy. Wei [46] proposed an occluded pedestrian detection method based on binocular vision. The Binocular introduced visual saliency prior information, which solves the problem of occlusion.

Algorithm based on recurrent neural network

Recurrent Neural Network (RNN) takes sequence data as input and recursively in the evolutionary direction of the sequence with all nodes are linked by a chain. Bidirectional RNN (Bi-RNN) and Long Short-Term Memory networks (LSTM) are common Recurrent Neural Network.

Stewart and Andriluka propose a model that is based on decoding an image into a set of people detections in crowded scenes, as it is shown in Fig. 8 [47]. A recurrent LSTM layer is used for sequence generation to train the model end-to-end with a new loss function that operates on sets of detections.

Comparison of typical experimental methods

Pedestrian databases

The MIT pedestrian database (MIT-CBCL Pedestrian Database) was created by the Massachusetts Institute of Technology. It contains 924 Pedestrian images (in PPM format with a width and height of 64×128). The images in the database contain both front and back perspectives. The images of USC Pedestrian Set are mostly from surveillance video, including three sets of data sets USC-A, USC-B, and USC-C. Daimler Pedestrian Detection Benchmark include

grayscale images. The database contains many images of occluded pedestrians. Caltech pedestrian database is a large-scale pedestrian database that has a relatively consistent pedestrian occlusion image with the actual situation in life. INRIA pedestrian database is the most widely used static pedestrian detection database having a clear picture. It has corresponding labeling files that are divided into a training set, test set, positive and negative samples. CUHK Occlusion Dataset, published by The Chinese University of Hong Kong, contains 1063 images of people. It has a large number of occluded pedestrian images. In addition, CUHK can release the “Person Re-identification Datasets” and “Square Dataset.” CUHK-Pre-D recorded 971 pedestrian samples from different perspectives. The CVC pedestrian database contains three subsets of cvc-01, cvc-02, and cvc-virtual, with each subset serving different tasks. The NICTA pedestrian database is a large static image pedestrian database, which is divided into a test set and a training set, containing 25,551 single person images and 5207 high-resolution non-pedestrian images. The images provided by the TUD pedestrian database are mainly convenient for calculating optical flow information. These databases are shown in Table 2, which are often used in pedestrian detection and tracking research.

Evaluation of multiple databases

Due to the use of widely varying evaluation protocols on multiple data makes direct comparisons difficult. An extensive evaluation of the state of the art can be performed in a unified framework, but it still has much room for improvement. In particular, detection is disappointing at low resolutions and for partially occluded pedestrians. Dollar calculated the frequency of pedestrian occlusion, which further divided pedestrians into four occlusion levels according

Table 2 Pedestrian databases

Databases	Institutions	Details	Download
MIT	MIT	924 pictures Includes front and back views	https://cbcl.mit.edu/software-datasets/PedestrianData.html
USC	Computer vision lab of USC	USC-A:313 standing pedestrians, USC-B: 271 pedestrian multi-angle occlusion, USC-C: 232 pedestrian multi-angle occlusion	https://sites.usc.edu/iris-cvlab/
Caltech	Caltech	10 h of 640 × 480 30 Hz video, Contains 350,000 rectangular boxes 2300 pedestrians	https://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/
INRIA	INRIA	Training set:614 positive samples '1218 negative samples; Test set:288 positive samples' 453 negative samples	https://pascal.inrialpes.fr/data/human/
Daimler	Daimler Lab	4800 pictures of people '5000 pictures of other objects', all 18 × 36 in size	https://www.science.uva.nl/research/isla/downloads/pedestrians/
CUHK	CUHK	1063 pictures of pedestrians including adhesion,occlusion	https://www.mmlab.ie.cuhk.edu.hk/datasets/cuhk_occlusion/index.html
CVC	–	Using Bumblebee2 stereo color vision system resolution of 640 × 480	https://www.cvc.uab.es/adas/index.php?section=other_datasets
TUD	Max Planck Institution	Positive samples include 1776 people, negative samples include 192 people	https://www.juhe.cn/market/product?%20id=10190
Cityperson	–	Five thousand images from 27 cities	https://www.cityscapes-dataset.net
NICTA	–	25,551 images with single person	https://www.nicta.com.au/category/research/computer-vision/tools/automap-datasets/

to the area occluded: full occlusion ($\geq 80\%$), heavy occlusion (35%–80%), partial occlusion (1–35%), never occlusion (0%). Most people think of comparing the performance of each window of an algorithm. Dalal suggests evaluating the detector by classifying a fixed-density sampling between a pedestrian-centered clipping window with an image without pedestrians.

These terms are used in the object detection: TP (True Positive) means to predict a positive sample to be a positive sample; FP (False Positive) means that the negative sample is predicted to be a positive sample; TN (True Negative) means to predict a negative sample to be a negative sample; FN (False Negative) means to predict a positive sample to be a negative sample. There are two indicators: Recall(R) and Miss rate(MR): $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$. $\text{MR} = 1 - \text{R}$.

In pedestrian detection, there are two indicators: MR-FPPI and MR^{-2} . FPPI: Assuming that the amount of error detec-

tion window in N images is k, then FPPI (false positive per image) is k/N . MR^{-2} : The value of MR^{-2} summarizes the performance of the detector using a log-average miss rate. The calculation method is the average miss rate under 9 FPPI values (range [0.01, 1.0]). The lower the score indicates better performance.

Tables 3 and 4 show the performance of several algorithms in INRIA and Caltech USA. Comparing Tables 3 and 4. Since most images in INRIA have no occlusion, the accuracy of HOG, HOG-LBP, MultiFtr + css and Franken is higher than that in Caltech USA partially occluded subset. The detection accuracy of these traditional algorithms is greatly affected when partial occlusion occurs. Especially, HOG algorithm does not specially deal with occlusion. This detector has the worst performance under occlusion. Although Franken has achieved good detection results on both database, there is still a certain gap in practical applications.

Table 3 Several traditional detection algorithms in INRIA

Method	Classifier	Pedestrian database	Log-average miss rate (%)
HOG	Linear SVM	INRIA	46
HOG-LBP	Linear SVM	INRIA	25.18
MultiFtr + css	Linear SVM	INRIA	23.93
Franken	2000 weak classifiers	INRIA	13.7

Table 4 Several traditional detection algorithms in Caltech USA partially occluded subset

Method	Classifier	Pedestrian database	Log-average miss rate (%)
HOG	Linear SVM	C-USA partially occluded subset	68.46
HOG-LBP	Linear SVM	C-USA partially occluded subset	64
MultiFtr + css	Linear SVM	C-USA partially occluded subset	61.46
Franken	2000 weak classifiers	C-USA partially occluded subset	40.45

According to Table 5, deep learning algorithms in different occlusion subsets of CityPerson perform differently [48–50]. MR^{-2} is used to compare the performance of deep learning detectors (lower score indicates better performance). In Table 5, the performance of these algorithms on reasonable subset and partial subset is similar except RPN + BF. This shows that these algorithms are capable of handling partial occlusion. RPN + BF is a high precision algorithm, but it does not deal with occlusion. So, its accuracy changes greatly when

Fig. 9 The performance on INRIA dataset [44] (The circle represents the traditional algorithm, and the triangle represents the deep learning algorithm)

occlusions occur. What is more, in the case of heavy occlusion, all algorithms accuracy will decline rapidly.

Figure 9 [44] compares the performance of deep learning algorithms and traditional algorithms on INRIA. (The circle represents traditional algorithms, and the triangle represents deep learning algorithms) It shows that the deep learning algorithm has more advantages and higher accuracy than that of the traditional algorithm. OR-CNN has the best performance. Figures 10, 11, 12 [39] reports the deep learning algorithms' and traditional algorithms' results on Caltech reasonable, partial occlusion, and heavy occlusion subsets, respectively. The main algorithms include DeepParts, HOG, MT-DPM, JointDeep, SDN, ACF + SDT, AlexNet and so on. In reasonable subset, the performance of deep learning algorithm is better than that of traditional algorithm. As the occlusion part increases, the gap between traditional algorithms and deep learning algorithms shrinks. Nevertheless, deep learning algorithms still perform better. The accuracy of the algorithms with special treatment for occlusion is less affected.

Table 5 Pedestrian detection results in City Persons

Method	Pedestrian database	Reasonable (MR^{-2})	Partial (MR^{-2})	Heavy (MR^{-2})
Adapted FasterR-CNN	CityPersons	12.8	—	—
Repulsion Loss	CityPersons	11.6	14.8	55.3
OR-CNN	CityPersons	11.0	13.7	51.3
Zhang et al	CityPersons	15.4	18.9	55.0
CAM-based Attention	CityPersons	13.61	—	46.17
RPN + BF-P1	CityPersons	10.1	18.9	58.9

Fig. 10 Average miss rate on reasonable subset of Caltech [39]

Fig. 11 Average miss rate on partial subset of Caltech

Fig. 12 Average miss rate on heavy occlusion subset of Caltech

Conclusion

In this paper, pedestrian detection methods under occlusion are reviewed. First, pedestrian detection algorithms based on traditional methods and deep learning are introduced. Second, for each class of methods, according to the different treatment of occlusion, the traditional methods are further divided into two categories, and the deep learning method is divided into three categories. The results show that the algorithm based on the traditional method that manually selects pedestrian features to train algorithm is time-consuming and less robust. The deep learning method has better performance

speed, which is more suitable for practical application and has a broad development prospect.

Although, pedestrian detection under occlusion has achieved an excellent recognition effect, there are still many problems to be solved in complex traffic situations or scenes with the massive human flow and it mainly includes:

1. Training data problem: In the case of a small amount of data, the current algorithms can not get a good detection effect. At present, most algorithms are trained in large data sets to fine-tune the trained models.
2. Robustness and speed problem. The detection accuracy and detection speed are always challenging to be considered in pedestrian detection technology. When the detection accuracy is guaranteed, the model needs to learn the characteristics of pedestrians thoroughly, which increase the amount of calculation and store more data that are inevitably lead to the slow detection speed and failure to meet the demand of real time. To ensure the detection speed, usually reducing the amount of calculation will lead to insufficient training. Therefore, it is significant to design an efficient algorithm with both detection accuracy and detection speed.
3. Long-term occlusion or heavy occlusion problem: From the comparison results of the algorithms in this paper, the pedestrian detection algorithm for occlusion has an excellent performance in the case of slight or partial occlusion. However, in the case of heavy occlusion or long-term occlusion, the accuracy will decline rapidly. Therefore, efforts are needed to solve the problem of long-term and severe occlusion.

Acknowledgements This work is supported in part by China National Textile and Apparel Council No.2018097, National Natural science Foundation of China 61902301 and Shaanxi Provincial Education Department 19JK036418JK0334. Thanks all reviewers.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Risse B, Mangan M, Del Pero L, Webb B (2017) Visual tracking of small animals in cluttered natural environments using a freely moving camera. In: IEEE international conference on computer vision workshops (ICCVW) Venic, pp 2840–2849
- Luo Y, Yin D, Wang A, Wu W (2018) Pedestrian tracking insurveillance video based on modified CNN. *Multimed Tools Appl* 77:24041–24058
- Brunetti A, Buongiorno D, Trotta GF, Bevilacqua V (2018) Computer vision and deep learning techniques for pedestrian detection and tracking: a survey. *Neuro Comput* 300:17–33
- Hou L, Wan W, Hwang JN, Muhammad R, Yang M, Han K (2017) Human tracking over camera networks: a review. *Eurasip J Adv Signal Process* 1:356–367
- Chang M-F, Lambert J, Sangkloy P, Singh J, Sławomir B, Hartnett A, Wang D, Carr P, Lucey S, Ramanan D, Hays J (2019) Argoverse: 3D tracking and forecasting with rich maps. *IEEE Conf Comput Vision Pattern Recognit*, pp 8748–8757
- Luo W, Yang B, Urtasun R (2018) Fast and furious: real-time end-to-end 3D detection, tracking and motion forecasting with a single convolutional net. *IEEE/CVF conference on computer vision and pattern recognition*. Salt Lake City, UT, pp 3569–3577
- Girao P, Asvadi A, Peixoto P, Nunes U (2016) 3D object tracking in driving environment: a short review and a benchmark dataset. In: IEEE intelligent transportation systems conference, Rio de Janeiro, pp 7–12
- Li C, Liang X, Lu Y, Zhao N, Tang J (2019) RGB-T object tracking: benchmark and baseline. *Pattern Recogn* 96(1):67–79
- Hoof HV, Zant TVD, Wiering M (2011) Adaptive visual face-tracking for an autonomous robot. In: Belgian/Netherlands artificial intelligence conference, Nov 3 2011–Nov 4. 2011, pp 25–37
- Robin C, Lacroix S (2016) Multi-robot target detection and tracking: taxonomy and survey. *Autonomous Robots* 40(4):729–760
- Dollár P, Wojek C, Schiele B, Perona P (2012) Pedestrian detection: an evaluation of the state of the art. *IEEE Trans Pattern Anal Mach Intell* 34(4):743–761
- Dollár P, Appel R, Kienle W (2012) Crossover cascades for frame-rate pedestrian detection. In: *Proceedings of the European conference on computer vision (ECCV)*, pp 645–659
- Hinton GE, Osindero S, Teh Y (2006) A fast learning algorithm for deep belief nets. *Neural Comput* 18(3):1527–1554
- Lienhart R, Maydt J (2002) An extended set of haar-like features for rapid object detection. In: *International conference on image processing*, Rochester, NY, USA, pp 1–1
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: *IEEE computer society conference on computer vision and pattern recognition*, San Diego, CA, USA, vol 1, pp 886–893
- Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. *European conference on computer vision*. Springer, Cham, pp 818–833
- Simonyan K, Zisserman A (2016) Very deep convolutional networks for large-scale image recognition. *Comput Sci* 25(1):140–156
- Redmon J, Divvala S, Girshick R et al (2016) You only look once: Unified, real-time object detection. In: *IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, 2016, pp 779–788
- Leibe B, Seemann E, Schiele B (2005) Pedestrian detection in crowded scenes. In: *IEEE computer society conference on computer vision and pattern recognition*, San Diego, CA, USA, 2005, vol 1, pp 878–885
- Mohan A, Papageorgiou C, Poggio T (2001) Example-based object detection in images by components. *IEEE Trans PAMI* 23(4):349–361
- Mikolajczyk K, Schmid C, Zisserman A (2004) Human detection based on a probabilistic assembly of robust part detector. *Eur Conf Comput Vis* 1:69–82
- Wu B, Nevatia R (2009) Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. *Int J Comput Vision* 82(2):185–204
- Wang X, Han X, Yan S (2009) An hog-lbp human detector with partial occlusion handling. In: *IEEE international conference on computer vision*, Kyoto, pp 32–39
- M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila (2010) Multi-cue pedestrian classification with partial occlusion handling. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, pp. 990–997
- Flores Calero J, Aldás M, Lázaro J, Gardel A, Onofa N, Quinga B (2019) Pedestrian Detection Under Partial Occlusion by using Logic Inference, HOG and SVM. *IEEE Latin America Transactions* 17(09):1552–1559
- Wojek C, Walk S, Roth S, Schiele B (2011) Monocular 3D scene understanding with explicit occlusion reasoning. *IEEE Computer Vision and Pattern Recognition*. CVPR Providence, RI 2011:1993–2000
- Mathias M, Benenson R, Timofte R, Van Gool L (2013) Handling occlusions with franken-classifiers. *IEEE Int Conf Comput Vis Sydney NSW* 2013:1505–1512
- Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. *IEEE Trans PAMI* 32(9):1627–1645
- DuX, El-Khamy M, Lee J, Davis L (2017) Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In: *IEEE winter conference on applications of computer vision (WACV)*, Santa Rosa, CA, pp 953–961
- Zhang S, Wen L, Bian X, Lei Z, Li SZ (2018) Single-shot refinement neural network for object detection. In: *IEEE/CVF conference on computer vision and pattern recognition*, Salt Lake City, UT, pp 4203–4212
- Ouyang W, Wang X (2012) A discriminative deep model for pedestrian detection with occlusion handling. In: *IEEE conference on computer vision and pattern recognition*, Providence, RI, pp 3258–3265
- Ouyang W, Wang X (2013) Joint deep learning for pedestrian detection. In: *IEEE international conference on computer vision*, Sydney, NSW, pp 2056–2063
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S (2016) SSD: single shot multibox detector. In: *European conference on computer vision lecture notes in computer science*, vol 9905. Springer, Cham, pp 2103–2112
- Fu C-Y, Liu W, Ranga A, Tyagi A, Berg AC (2016) DSSD: Deconvolutional single shot detector. *Sci Chin Inf Sci* 63(2):113–120
- Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, Las Vegas, NV, pp 779–788
- Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. In: *IEEE conference on computer vision and pattern recognition*, Honolulu, HI, pp 6517–6525
- Ouyang W, Wang X, Zeng X, Qiu S, Luo P, Tian Y, Li H, Yang S, Wang Z, Loy C-C, Tang X (2015) Deepid-net: deformable deep convolutional neural networks for object detection. In: *IEEE Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, pp 2403–2412
- Ouyang W, Zhou H, Li H (2018) Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *IEEE Trans Pattern Anal Mach Intell* 40(8):1874–1887

39. Tian Y, Luo P, Wang X, Tang X (2015) Deep learning strong parts for pedestrian detection. In: IEEE international conference on computer vision (ICCV), Santiago, pp 1904–1912
40. Shanshan Z, Jian Y, Bernt S (2018) Occluded pedestrian detection through guided attention in CNNs. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, Salt Lake, UT, pp 6995–7003
41. Zou T, Yang S, Zhang YY, Ye M (2020) Attention guided neural network models for occluded pedestrian detection. *Pattern Recogn Lett* 131(1):91–97
42. Zhou C, Yuan J (2017) Multi-label learning of part detectors for heavily occluded pedestrian detection. In: IEEE international conference on computer vision (ICCV), Venice, pp 3506–3515
43. Wang X, Xiao T, Jiang Y, Shao S, Sun J, Shen C (2017) Repulsion loss: detecting pedestrians in a crowd. *IEEE/CVF Conf Comput Visi Pattern Recognit Salt Lake, UT 2018*:7774–7783
44. Shifeng Z, Wen L, Bian X, Lei Z, Li SZ (2018) Occlusion-aware R-CNN: detecting pedestrians in a crowd. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) *Computer Vision -ECCV 2018*. ECCV, pp 6885–6997
45. Sumi A, Santha T (2019) Frame level difference (FLD) features to detect partially occluded pedestrian for ADAS. *J Sci Ind Res* 78(12):831–836
46. Wei W, Cheng L, Xia Y (2019) Occluded pedestrian detection based on depth vision significance in biomimetic binocular. *IEEE Sens J* 19:11469–11474
47. Stewart R, Andriluka M (2016) End-to-end people detection in crowded scenes. In: IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, 2016, pp 2325–2333
48. Zhou C, Yuan J (2016) Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection. In: Lai SH, Lepetit V, Nishino K, Sato Y (eds) *Computer Vision. Lecture Notes in Computer Science*, vol 10112. pp 1146–1160
49. Zhang S, Benenson R, Schiele B (2017) Citypersons: a diverse dataset for pedestrian detection. In: IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, pp 4457–4465
50. Zhou C, Yuan J (2019) Multi-label learning of part detectors for occluded pedestrian detection. *Pattern Recogn* 86(2):99–111

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.