Computer Networks 168 (2020) 107036

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Faster R-CNN for multi-class fruit detection using a robotic vision system

Shaohua Wan[a,*], Sotirios Goudos[b]

[a] School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073 China
[b] Department of Physics, Aristotle University of Thessaloniki, Thessaloniki GR-54124, Greece

ARTICLE INFO

Article history:
Received 30 May 2019
Revised 7 November 2019
Accepted 26 November 2019
Available online 4 December 2019

Keywords:
Deep learning
Faster R-CNN
Multi-class fruit detection
Image region selection

ABSTRACT

An accurate and real-time image based multi-class fruit detection system is important for facilitating higher level smart farm tasks such as yield mapping and robotic harvesting. Robotic harvesting can reduce the costs of labour and increase fruit quality. This paper proposes a deep learning framework for multi-class fruits detection based on improved Faster R-CNN. The proposed framework includes fruits image library creation, data argumentation, improved Faster RCNN model generation, and performance evaluation. This work is a pioneer to create a multi-labeled and knowledge-based outdoor orchard image library using 4000 images in the real world. Also, improvement of the convolutional and pooling layers is achieved to have a more accurate and faster detection. The test results show the proposed algorithm has achieved higher detecting accuracy and lower processing time than the traditional detectors, which has excellent potential to build an autonomous and real-time harvesting or yield mapping/estimation system.

1. Introduction

As one of the object recognition technologies, fruit classification isnt only mainly applied to fruit quality detection, sorting by classifications, maturity identification, defect detection and robot picking, but also widely used in the field of intelligent agriculture. The technology of fruit classification principally integrates the knowledge of multi fields including image processing and pattern recognition to obtain a feature set of fruit, through training and learning of which to classify fruits. In the real environment where the variety of fruits and upgraded types are bred increasingly, how to quickly recognize various fruits by robots and improve the precision of fruit classification to enhance peoples cognition level of multi-species becomes one of the main targets of researching on algorithm of fruit classification. At present, researchers have achieved a lot in the analysis and processing of two-dimensional images, namely color images of fruits, however, it is limited that the precision of recognition is always inferior to the actual application due to the influence of the conditions of light, coverage, visual angles, scale variation and environmental changes. In recent years, researchers have introduced technologies of 3D reconstruction images of fruit and RGB-D depth images for analyzing and processing, in order to develop application algorithms of improvement of fruit recognition precision.

Most researches on recognition of fruit classification focus on a certain category, a few of which concentrate on multi-fruits classification. The recognition of multi-fruits classification also has extensive value of practical application. For example, the recognition technology of multi-fruit images is used in self-service of buying fruits in supermarkets of developed countries. And in production line, it can also reduce errors caused by manual picking and improve production efficiency. Moreover, in intelligent agriculture, the classification recognition of fruit images can achieve precise breeding of fruit trees in multi-varieties mixed orchards and automatic picking of fruits. The non-destructive monitoring during fruit growth plays an important role in the production. The real-time and precise monitoring of fruit growth contributes to the fine management of trees, thereby increase the yield and quality of fruits. First, it is necessary to know the growth of fruits in different periods, which helps people analyze the relationship between fruit growth and the environment, whereupon guide the work of fertilization, irrigation, pest control, and raising fruit quality substantially. Secondly, the shapes of fruits have changes with the growth cycle, and these changes are noticeable in images. Therefore, it is indispensable and urgent to research on recognition and monitoring of fruit growth state by taking advantages of computer vision and image processing technologies. In production

* Corresponding author.
E-mail addresses: shaohua.wan@ieee.org (S. Wan), sgoudo@physics.auth.gr (S. Goudos).

https://doi.org/10.1016/j.comnet.2019.107036
1389-1286/

of fruit, picking is an important process demanding a large labor cost, as the fastening pace of mechanized production, decreasing people in rural areas specialize in agricultural production. As a result, the shortage of workforce will affect fruit picking seriously. It can't meet the requirements of modern agricultural production by manual picking fruits alone, not to speak of the risk in manual picking. With the rapid development of artificial intelligence and production technologies, it is possible to replace the manual fruit picking by machines. At present, researchers have developed a variety of robots for picking in different ways, but, these robots working is based on the premise that they can precisely and quickly recognize the fruits for picking. The precision and efficiency of the recognition system mostly determine the efficiency of robot picking. Although the traditional recognition method of fruit images performs great, it still cannot meet the requirements of commercial applications. Therefore, it is necessary to select a more suitable algorithm of fruit recognition to enhance the precision and efficiency of fruit recognition.

Mechanical visual technology can make picking robot position the target precisely. So, the development of picking robots with the visual function has great application value and strong practical significance for improving agricultural productivity. In recent years, Convolutional Neural Networks have shown great advantages in target detection, including two categories. One is a detection method generated from regions, by which CNN classifies the generated regions which may include the picking targets [1–6]. The other is regression-based method which uses CNN to process the whole image for target positioning and categorizing [7–10]. It is faster than the former. In order to compensate low precision in the detection of the later methods, researchers have used deep CNNs such as the net of He [11] and Szegedy [12] to improve precision. Nevertheless, the deep CNN often leads to higher computational complexity and data loss after multi-layers transmission [13–16]. The deep learning method is characterized by high precision and fast calculation speed, based on which, the paper takes orchard fruit as the main research object and uses the Faster RCNN method to recognize fruits of different sizes. Furthermore, in considering different situations of the natural environment, it selects training samples including green apples and mangoes with different sizes under different light conditions to design a visual detection method for green apples and mangoes in trees under natural environment. Eventually, a technical support is provided for the research of visual detection of green fruits in agricultural production.

Therefore, this paper proposes an improved Faster R-CNN for multiclass fruit detection using a deep learning framework to achieve higher efficiency, effectiveness, and reliability in outdoor orchard environments. The main contributions of this paper are highlighted as:

1. Establish the first self-learning based fruit image library with automatically tuning the parameters during the training process;
2. Propose data augmentation methods to perform detection on high resolution images;
3. Optimize the structure of the existing Faster R-CNN model in both convolutional layer and pooling layer.

The remainder of this paper consists of the following. Section 2 introduces related work and the background. Section 3 presents Faster Region-based Convolutional Neural Network (R-CNN) for multi-class fruit detection. Section 4 shows evaluation data and performance metrics. We demonstrate the experimental results in Section 5. Conclusions and future work are drawn in Section 6.

2. Related work

Fruit picking operation is a labor intensive work taking the most cost of time and labor in agricultural production, thereby it is urgent to realize automatic fruit picking to improve labor efficiency. With the rapid development of artificial intelligence and production technologies, it is possible to replace the manual fruit picking by machines. At present, researchers have developed a variety of robots for packing in different ways, but, these robots' working is based on the premise that they can precisely and quickly recognize the fruits for picking. The precision and efficiency of the recognition system mostly determine the efficiency of robot picking. At present, research on fruit detection has made great progress. Kapach et al [17] outlined the visual department of fruit picking robots, comprehensively expounding the advantages and limitations of various methods in the current field. Most of the traditional recognition method of fruit images involves the color of the target [18], while others use a combination of features such as color, texture, shape, etc. to achieve the target detection of fruit, but these methods are based on research in certain scenarios. They have poor robustness, low recognition efficiency, unable to meet the working needs of picking robots in various complex scenarios. Hence, we need to explore an algorithm of fruit recognition with the better precision and efficiency of fruit recognition.

Compared with conventional methods, the CNN has been showing great advantages in the field of target detection in recent years. The CNN make it possible to recognize fruits in complex situations due to their deep extraction of high-dimensional features of the targets. There are two methods. One is a method based on regional suggestions, which is represented by algorithms of RCNN, Fast RCNN and Faster RCNN [1,3,4,19]. The kernel idea is obtaining suggested regions first and then classifying them in the region. It is also called two-stage target detection. The other is a method without suggested regions. Typical algorithms are SSD [8] and YOLO [7]. The kernel idea is using a single convolutional network to detect the position of the target and its properties on the basis of the whole image. It is also called one-stage target detection. In recognizing immature mangoes, the YOLOv2 network maintains the precision and the generalization, simultaneously increase the detection rate, while the recall rate decreases in detecting regions with a big density of apple. It is easy to recognize a cluster of apples as one.

Multi-class fruit detection has been explored by many researchers in smart farming, across a variety of orchard types for the purposes of autonomous harvesting or yield mapping/estimation [17,20–25]. Bargoti and Underwood [20] proposed Tiled Faster R-CNN to design a trained model over large fruit images outside orchards, with an F1-score of > 0.9 attained for mangoes and apples. Parallel to this work, Sa et al. [22] presented the DeepFruits approach. They employed the Faster RCNN (FR-CNN) object detection method which used deep neural networks to jointly perform region proposal and classification of the proposed regions. Sa et al. attained up to a 0.83 F1 score with multi-class farm dataset, applying this approach to multi-class fruits, such as Sweet pepper, Rock melon, Apple, Avocado, Mango and Orange, and showed that such a method could be speedily trained and deployed. In [25–30], audio/video/image has been used with success, and demonstrate a better performance outside smart farming automation. Our work follows the same approach and proves the use of Faster R-CNN to realize the identification of multi-class fruits in orchards environments, speed up the detection without affecting the accuracy, and carry out network identification tests under different shooting time, growth stages, light and other interference scenarios, with a view to detect network performance and experiment with CPU and GPU configurations to verify real-time requirements.

3. Methods

This section proposes the Faster R-CNN approach for multi-class fruit detection in orchards and introduces more details on an improved Faster RCNN method, which are used within the experimental studies conducted in Section 5.

3.1. An improved faster R-CNN based model generation

Fruit detection framework based on improved Faster R-CNN, shown in Fig. 1, uses fruits images in the training sets as inputs and outputs classification results and bounding box coordinates of electrical components in the images.

A CNN model VGG-16 is used for the feature extraction process. This model includes 13 convolutional layers, 13 ReLu layers and four pooling layer. The inputs come from the original inspection images. Assume the lth layer is a convolutional layer, the output feature vector of this layer is:

$$x_j^l = f_l(\Sigma_{i\in M_j} x_i^{l-1} * k_{ij}^l + b_j^l) \tag{1}$$

where x_j is the ith output of this layer, k_{ij}^l is the convolutional kernel, $*$ means the convolution multiplication, b_j^l is the bias of this layer and $f_l(g)$ is the ReLu function. Assume the Mth layer is a pooling layer, the output feature vector of this layer is:

$$x_j^m = f_2(\beta_j^m S(x_j^{m-1}) + b_j^m) \tag{2}$$

where β_j is the connecting coefficient, x_j^{m-1} is the input for this pooling layer, $S(g)$ is the summation of the whole matrix, b_j^m is the bias, and $f_2(g)$ is the softmax function.

An over-fitting problem may occur in the training process when there are not much training samples. To further improve the accuracy of the detection and balance the model complexity and data volume, a regularization method is used to decay weights for some high dimensional parameters. Two loss functions are added to optimize the convolutional and pooling layers. The parameters can be adjusted automatically according to the different shooting angles, which guarantees the size of each convolutional layer and the kernel parameters are within a reasonable range. The loss function in the convolutional layer is defined as:

$$L_{cv}(W, x^k) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(y_k - W * x_{ik}^2 + \lambda \sum_{j=1}^{n} w_j^2\right)\right] \tag{3}$$

where $W = [\omega_1, \omega_2, \ldots, \omega_n]^T$ are the parameters for the kernel, n is the number of the kernels, $*$ means the convolution multiplication, $x^{(k)} = [x_{1k}, x_{2k}, \ldots, x_{mk}]^T$ is the kth input, m is the dimension of the database, y_k is the actual label of the kth sample, and λ is the regularization penalty factor. In the proposed loss function, a likelihood function and a regularization function. The outputs are as close as the inputs through these two functions.

The loss function in the pooling layer is defined as:

$$L_{pl}(W, x^k) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(y_k - U * x_{ik}^2 + \lambda \sum_{j=1}^{n} \mu_j^2\right)\right] \tag{4}$$

where the first expression is showing the reconstruction loss and the second expression is presenting the model complexity. $U = [\mu_1, \mu_2, \ldots, \mu_t]$ are the pooling parameters and t is the total number of the parameters. λ is usually determined by expert knowledge, and it will be chosen based on testing results. Then a convolutional feature map can be obtained for each inspection image.

An RPN takes an inspection image with any size as input and provides a set of rectangular region proposals as output. To obtain region proposals, a small networkâa 3 × 3 spatial window of the convolutional feature mapâis slid over the convolutional feature map output by the last shared convolutional layer. Every sliding window is mapped to a lower dimensional feature and becomes an intermediate layer. To keep as much information of the input feature vector as possible, the dimension of the feature maps are reduced and normalized through an ROI pooling layer. Particular conversion vectors including a 4 × 4, a 2 × 2 and a 1 × 1 vector are utilized for the pooling process. First, 16 regions are obtained using the 4 × 4 vector. Max pooling is done in each region. Next, the 2 × 2 and 1 × 1 vectors are applied, and the same operation is done. As a result, the reduced dimension of any proposed region with different sizes is set to be 21. Then this layer is fed into two siblings fully connected layers-a box classification layer and a box regression layer. At each sliding-window location, multiple region proposals are simultaneously predicted, where the number of maximum possible proposals for each location is denoted as k. In this paper $k = 9$ anchors at each sliding position are utilized. The classifier and regressor work as follows:

1. Classifier: The probability of the region proposal contains an object is calculated in this layer. After the feature map is fully looked through, the probability P_i of each element in the

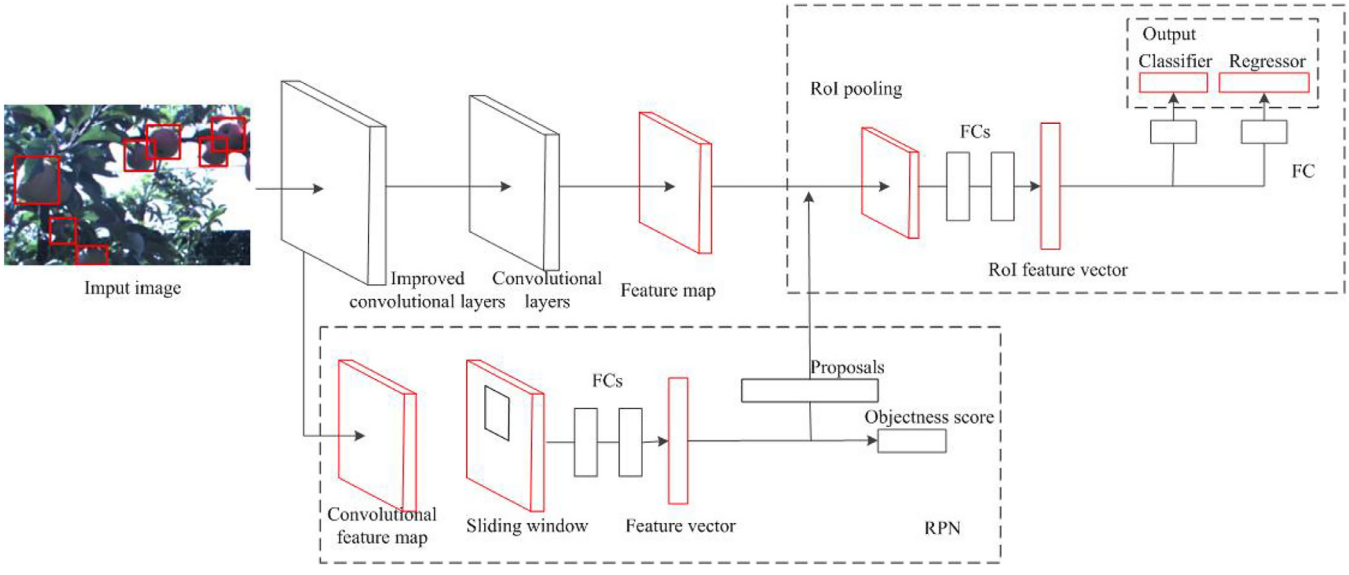


Fig. 1. The Improved Faster R-CNN Network.

feature map i contains the target object is calculated for using softmax function. The regions that have the top 300 P_i in the total rankings are the region proposals.

2. Regressor: In the regressor, Intersection over Union (IoU) is used as the index to measure the accuracy of the bounding box, and therefore to detect the coordinates (x, y) of the center point of the anchor box and the width w and length h of this box. IoU is calculated as:

$$IoU = \frac{A \cap B}{A \cup B} \tag{5}$$

where A and B are the area of two region proposals. Assume (x, y) are the coordinates of the center point in the proposed region, and (w, h) are the corresponding width and height of this region. Then a proposed region can be described using a four dimensional index $(x, y, w, h).(P_x, P_y, P_w, P_h), (G'_x, G'_y, G'_w, G'_h) - and(G_x, G_y, G_w, G_h)$ are representing the anchor box, predicted box and ground-truth box respectively. The window of IoU can adjust the edges of the original proposed object region in a regression process.

The error between the original proposed region and actual region is $t_k = (t_x, t_y, t_w, t_h)$, where:

$$t_x = \frac{G_x - P_x}{P_w}, t_y = \frac{G_y - P_y}{P_h} \tag{6}$$

$$t_w = log\left(\frac{G_w}{P_w}\right), t_h = log\left(\frac{G_h}{P_h}\right) \tag{7}$$

Therefore, the objective function can be written as:

$$L_e = \sum_{i=1}^{N} (t_K^i - \omega_K^T \phi(P_i))^2 \tag{8}$$

where $\phi(P_i)$ is the input feature vector, ω_{K^T} is are the coefficients obtained from training, and N is the total number of samples. The objective function can be solved by the least square method. Therefore the classifier outputs $2k$ scores that estimate the probability of object or background for each proposal, while the regressor outputs $4k$ outputs encoding the coordinates of k boxes. The k region proposals are associated with k reference boxes, denoting as anchor boxes. Note that, an RPN can be obtained using the same CNN model without training a separate model, which hugely reduced the processing speed.

4. Evaluation data and performance metrics

4.1. Data acquisition

The fruits data tested in this paper is composed of three fruit varieties: apples, oranges and mangoes. The data comes from Fruits 360 dataset [31]. All images were 100 × 100 pixels and has white background. The fruits were filmed while rotating around a fixed axis for either two or three different axises per fruit. Then images are acquired and some of them are selected as training set, testing set and validation set. The labelled dataset for each fruit is split into training, validation and testing sets (see Table 1).

Table 1
Dataset configuration.

Fruit	Image size	Train	Validation	Test
Apple	100 × 100	492	164	164
Mango	100 × 100	490	166	166
Orange	100 × 100	479	160	160

4.2. Data augmentation

Enhancement of the sample images can improve the quality and diversity of the sample, which contribute to improve the precision of CNN detection. For example, the “Stern” increases samples with brightness changes to improve detection precision. When the orchard is under natural lighting, especially the strong light, the fruits external shadow formed by partial coverage of mangos or photographing against the light direction make the big discrepancy of the color between regions under normal lighting and diffuse regions. It affects the quality of the images of fruit samples, thereby affects the detection of the model. The paper enhances the fruit image and reduces the light influence on the quality by technology of Adaptive histogram equalization, which is equivalent to adjusting the brightness of the images and increasing the diversity of lighting conditions of the samples. Therefore, this method is used to improve the image quality and increase samples. In the paper, considering that most of fruits are on the trees in a vertical hanging state and some are in slant hanging state with different angles caused by branches conditions and mutual coverage of fruits, the samples are horizontally mirrored and rotated 10 and 20, and the rotated images are intercepted from the center. If there is any imperfectness or even complete loss in target of the images’ edge after rotation, the tagging will be abandoned.

4.3. Evaluation measures

In this paper, the objective evaluation criteria are used to evaluate the fruit recognition system. It uses the precision, recall, and IoU to evaluate the trained model, and find the appropriate threshold for the model, finally select the right target by the confidence coefficient of the model prediction. In the formula, TP is the true number of positive samples, FP is the number of false positive samples, FN is the number of false negative samples, C is the number of categories, N is the number of reference thresholds, k is the threshold, P(k) is the precision rate, and R(k) is the recall rate. The AP is the area under the Precision-recall curve. Generally speaking, the better the classifier is, the higher the AP value is. mAP is the average of multiple categories of APs. The meaning of mean is to average the APs of each category to get the value of mAP. The value of mAP must be in the range of [0, 1], which is the bigger, the better. This indicator is the most important one of the target detection algorithms.

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$mAP = \frac{1}{C} \sum_{k=i}^{N} P(K) \Delta R(k) \tag{11}$$

5. Experimental results and analysis

5.1. Experimental setup

All training and testing in this paper is performed on the same laptop with the main configuration of PC Intel Core(TM) i7-7700HQ CPU@2.80 GHz, 6GB GPU GeForce GTX 1060 and 16GB random access memory. All programs are written in C++ and call CUDA, OPENCV base and run under Windows10 system. The CPU and GPU are used separately in training and testing the model. The selection of learning rates and momentums directly influences the training speed and results of the detection network. In this paper, some common learning rates and momentums are selected as candidate

Table 2
Calculation results using different penalty factors.

λ	Iteration	N	mAP(%)	Speed(ms/image)
0	5000	1000	78.87	98
		300	79.65	89
10	5000	1000	80.36	92
		300	81.23	82
20	5000	1000	82.63	83
		300	82.96	78
50	5000	1000	82.87	80
		300	83.10	70
100	5000	1000	83.67	77
		300	84.57	66
200	5000	1000	84.12	70
		300	83.79	61
1000	5000	1000	83.16	63
		300	82.76	57

Table 3
Results from different detection and recognition methods.

Algorithm	mAP(%)	Speed(ms/image)	Manually tune parameters
YOLOv3	84.89	39	No
Faster-CNN	82.56	69	No
Improved Faster-CNN	86.41	47	No

values. Furthermore, it confirms better learning rates and momentums by the trials. The trials selected 0.1, 0.01, and 0.001 as candidate values of the initial learning rate, 0.5, 0.9, and 0.99 as candidate values of the momentum coefficient.

5.2. Parameter determination and selected region proposals

To exam the impact of penalty factor in Eq. (3) λ on the detection accuracy, tests are done using the improved Faster R-CNN model. Table 2 shows the mAP and processing speed for each condition. Here N is the number of images used for parameter test. When $\lambda = 200$, mAP is the highest, and the processing speed is satisfying. Therefore the penalty factor is chosen as $\lambda = 200$. The fruit components are precisely selected as the region proposals, which shows the effectiveness of the obtained RPN network.

5.3. Results and discussion

The features are exactly extracted, and the target components are accurately detected. They can be used for any images without manually tuning the parameters. Meanwhile, the performance of these methods is tested. The results are shown in Table 3. From Table 3, it can be seen that the digital image processing method and machine learning method can detect the multiclass fruits in the image with a less promising mAP. Also, the parameters of all these methods need to tuned manually, which makes them not good fits for big data analysis. CNN-based deep learning algorithms have better precision and higher image processing speed. Next, comparison results are shown to explore the performance of the proposed improved Faster R-CNN among all the CNNbased deep learning algorithms.

The proposed improved Faster R-CNN based multiclass fruits detection method is tested to show its advantages over the other existing methods. Comparing the testing results with the ground truth information, it is observed that YOLO has the worst performance among all the six methods. Improved Faster R-CNN, YOLOv2 and YOLOv3 are showing better detecting performance than Fast R-CNN and YOLO. YOLOv2, YOLOv3, and the proposed improved Faster R-CNN algorithms have the best detection results. They can accurately detect those components even with quite small sizes. Furthermore, the overall calculation performance of these six algorithms is shown in Table 4. The improved Faster R-CNN has higher detection precision. The improved Faster R-CNN has better recognition effect for apples, mangos and oranges; YOLO algorithm cannot be used for small-sized elements. Note that, Fast R-CNN has the good detection accuracy, but the processing speed is the lowest among all the CNN based algorithms. Also, it needs a more substantial selection space from the test, which makes it unable to meet the calculation requirement for future online detection. The proposed improved Faster R-CNN has the highest detection accuracy compared to the other algorithms due to the accurate and proper creation of the inspection image library and the layer optimizations. And also the parameters are better trained which helps to increase the detection speed. This proposed method can guarantee an accurate and fast detection for multiclass fruits in Orchards. The experimental results demonstrate that the improved Faster R-CNN algorithm is not only suitable for the detection of single fruit, but also for the high detection accuracy of the multi-class fruits, and it can be used in the orchard picking robot.

Table 4
Results from different detection and recognition methods.

Algorithm	mAP(%)			mAP	Speed
	Apple	Mango	Orange	(%)	(ms/image)
YOLO	78.79	70.94	60.69	70.14	60
Fast R-CNN	78.76	79.83	76.56	78.38	280
Faster R-CNN	86.87	89.36	87.39	87.87	90
YOLOv2	90.67	88.12	87.67	88.82	46
YOLOv3	91.89	89.52	88.70	90.03	40
Improved Faster R-CNN	92.51	88.94	90.73	90.72	58

6. Conclusions

This paper proposes a deep learning framework for multi-class fruits detection based on improved Faster R-CNN. The proposed framework includes fruits image library creation, data argumentation, improved Faster RCNN model generation, and performance evaluation. The key contributions are creating fruit image library and optimizing the structure of the convolutional and pooling layers in the model. The proposed improved Faster R-CNN based framework showed above 91% mAP including apples, mango and orange detection. The image processing speed is also increased compared to the existing algorithms, which makes the proposed method more general to be applied for actual scenes. The proposed framework has significant benefit infield automation in agriculture. It also achieves higher efficiency, effectiveness, and reliability within the smart farming environment. This work is the potential for this robotic vision system to perform not only fruit detection but also quality estimation. The future work includes creating a more extensive outdoor orchard image library to consider more multi-class fruits with various types of component defect and building an autonomous harvesting or yield mapping/estimation system.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared toinfluence the work reported in this paper.

Acknowledgments

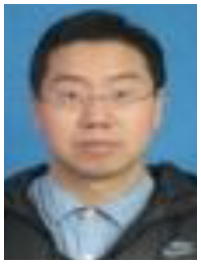
This work was supported in part by the Fundamental Research Funds for the Central Universities of China under Grant 2722019PY052 and by the open project from the State Key Laboratory for Novel Software Technology, Nanjing University, under Grant No. KFKT2019B17.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.comnet.2019.107036.

References

[1] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[2] Z. Gao, H.-Z. Xuan, H. Zhang, S. Wan, K.-K.R. Choo, Adaptive fusion and category-level dictionary learning model for multi-view human action recognition, IEEE. Internet. Things J. (2019).
[3] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[4] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing systems, 2015, pp. 91–99.
[5] S. Ding, S. Qu, Y. Xi, A.K. Sangaiah, S. Wan, Image caption generation with high-level image features, Pattern Recognit. Lett. (2019).
[6] Z. Gao, D. Wang, S. Wan, H. Zhang, Y. Wang, Cognitive-inspired class-statistic matching with triple-constrain for camera free 3D object retrieval, Future Gener. Comput. Syst. 94 (2019) 641–653.
[7] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.
[9] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.
[10] J. Redmon, A. Farhadi, YOLOv3: an incremental improvement, arXiv:1804.02767(2018).
[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[13] N. Lv, C. Chen, T. Qiu, A.K. Sangaiah, Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in SAR images, IEEE Trans. Ind. Inf. 14 (12) (2018) 5530–5538.
[14] S. Wan, Z. Gu, Q. Ni, Cognitive computing and wireless communications on the edge for healthcare service robots, Comput. Commun. (2019).
[15] W. Li, X. Liu, J. Liu, P. Chen, S. Wan, X. Cui, On improving the accuracy with auto-encoder on conjunctivitis, Appl. Soft Comput. 81 (2019) 105489.
[16] S. Ding, S. Qu, Y. Xi, S. Wan, Stimulus-driven and concept-driven analysis for image caption generation, Neurocomputing (2019).
[17] K. Kapach, E. Barnea, R. Mairon, Y. Edan, O. Ben-Shahar, Computer vision for fruit harvesting robots–state of the art and challenges ahead, Int. J. Comput. Vis. Robot 3 (1/2) (2012) 4–34.
[18] C. Zhao, W.S. Lee, D. He, Immature green citrus detection based on colour feature and sum of absolute transformed difference (SATD) using colour images in the citrus grove, Comput. Electron. Agric. 124 (2016) 243–253.
[19] L. Wang, H. Zhen, X. Fang, S. Wan, W. Ding, Y. Guo, A unified two-parallel-branch deep neural network for joint gland contour and segmentation learning, Future Gener. Comput. Syst. (2019).
[20] S. Bargoti, J. Underwood, Deep fruit detection in orchards, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 3626–3633.
[21] M. Halstead, C. McCool, S. Denman, T. Perez, C. Fookes, Fruit quantity and quality estimation using a robotic vision system, arXiv:1801.05560(2018).
[22] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, C. McCool, DeepFruits: a fruit detection system using deep neural networks, Sensors 16 (8) (2016) 1222.
[23] S. Bargoti, J. Underwood, Image segmentation for fruit detection and yield estimation in apple orchards, J. Field Rob. (2016).
[24] S. Nuske, K. Wilshusen, S. Achar, L. Yoder, S. Narasimhan, S. Singh, Automated visual yield estimation in vineyards, J. Field Rob. 31 (5) (2014) 837–860.
[25] I. Lenz, H. Lee, A. Saxena, Deep learning for detecting robotic grasps, Int. J. Rob. Res. 34 (4–5) (2015) 705–724.
[26] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 689–696.
[27] A. Eitel, J.T. Springenberg, L. Spinello, M. Riedmiller, W. Burgard, Multimodal deep learning for robust RGB-D object recognition, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2015, pp. 681–687.
[28] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, arXiv:1502.03167(2015).
[29] Y. Xi, Y. Zhang, S. Ding, S. Wan, Visual question answering model based on visual relationship detection, Signal Processing: Image Communication 80 (2020), 115648.
[30] Y. Zhao, H. Li, S. Wan, A. Sekuboyina, X. Hu, G. Tetteh, M. Piraud, B. Menze, Knowledge-aided convolutional neural network for small organ segmentation, IEEE journal of biomedical and health informatics 23 (4) (2019) 1363–1373.
[31] H. Mureşan, M. Oltean, Fruit recognition from images using deep learning, Acta Univ. Sapientiae Informatica 10 (1) (2018) 26–42.



Shaohua Wan (SM'19) received the joint Ph.D. degree from the School of Computer, Wuhan University and the Department of Electrical Engineering and Computer Science, Northwestern University, USA in 2010. Since 2015, he has been holding a post-doctoral position at the State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology. From 2016 to 2017, he was a visiting professor at the Department of Electrical and Computer Engineering, Technical University of Munich, Germany. He is currently an Associate Professor with the School of Information and Safety Engineering, Zhongnan University of Economics and Law. His main research interests include deep learning for Internet of Things and Cyber-Physical Systems. He is an author of over 60 peer-reviewed research papers and books. He is a senior member of IEEE.

Sotirios K. Goudos received the B.Sc. degree in Physics in 1991 and the M.Sc. of Postgraduate Studies in Electronics in 1994 both from the Aristotle University of Thessaloniki. In 2001, he received the Ph.D. degree in Physics from the Aristotle University of Thessaloniki and in 2005 the Master in Information Systems from the University of Macedonia, Greece. In 2011, he obtained the Diploma degree in Electrical and Computer Engineering from the Aristotle University of Thessaloniki. He joined the Department of Physics, Aristotle University of Thessaloniki in 2013, where he is currently an Assistant Professor. Dr. Goudos is the director of the ELEDIA@AUTH lab member of the ELEDIA Research Center Network. His research interests include antenna and microwave structures design, evolutionary algorithms, machine learning, wireless communications, and semantic web technologies. Dr. Goudos is currently serving as Associate Editor for IEEE ACCESS. He is also member of the Editorial Board of the International Journal of Antennas and Propagation (IJAP), the International Journal of Energy Optimization and Engineering, the EURASIP Journal on Wireless Communications and Networking and the International Journal on Advances on Intelligent Systems. Dr. Goudos was the Lead Guest Editor in the 2016 and 2017 Special Issues of the IJAP with topic "Evolutionary Algorithms Applied to Antennas and Propagation: Emerging Trends and Applications". He was the Editor of the book "Microwave Systems and Applications", InTech publishers, 2017. Dr. Goudos has served as the Technical Program Chair in the International Conference on Modern Circuits and Systems Technologies (MOCAST). He was sub-committee chair in the Asian-Pacific Microwave Conference (APMC 2017) in the track of Smart and reconfigurable antennas. He has also served as a member of the Technical Program Committees in several IEEE and non-IEEE conferences. Dr. Goudos is a member of the IEEE (senior member), the IEICE, the Greek Physics Society, the Technical Chamber of Greece, and the Greek Computer Society.