IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 18, 2020, accepted June 11, 2020, date of publication June 16, 2020, date of current version June 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3002882

A Shallow Convolutional Neural Network for Apple Classification

JINQUAN LI, (Member, IEEE), SHANSHAN XIE, ZHE CHEN, HONGWEN LIU, JIA KANG, ZIXUAN FAN, AND WENJIE LI

School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding authors: Jinquan Li (lijinquan@bupt.edu.cn) and Zhe Chen (zhechen95@qq.com)

ABSTRACT In the automatic apple sorting task, it is necessary to automatically classify certain apple species. A shallow convolutional neural network (CNN) architecture is proposed for this purpose. After collecting a certain number of apple images and labelling them, training data is obtained through a series of data augmentation operations, and then training and parameter optimization are carried out through the Caffe framework. The feasibility of the method is verified by experiments which are divided into two cases. In the case of no occlusion, the classification accuracy of apple images reaches approximately 92% in our test set. Besides, block voting is used to aid the proposed method and a good result can be achieved in our test set in the case of part occlusion caused by branches and leaves, rotten spots, and other kinds of apples. The proposed shallow network is characterized by a small number of parameters and shows resistance to overfitting with a limited dataset. Such a network presents an alternative for classification related tasks in smart visual Internet of Things and brings attention to reducing the complexity of deep neural networks while maintaining their strength.

INDEX TERMS Image classification, CNN, overfitting, smart visual Internet of Things.

I. INTRODUCTION

In the agricultural field, various agricultural applications and their degree of automation keep increasing continuously due to progress made in image recognition and classification technology. In [1], image technology is applied to weed identification in precision agriculture. Weed types are classified according to online weeds and chemical applications at specific locations. In [2], neural networks and maximum likelihood classifiers are applied. A single weed and crop mixture is classified to obtain the position and density of weeds in the crop. In [3], Gabor wavelet and gradient field distribution are combined to extract a new set of feature vectors for weed classification according to the direction texture features of weeds. In [4], machine learning, image processing, and classification-based methods are used to identify and detect diseases on agricultural products. In [5], gray-scale co-occurrence matrices and principal components are used to extract leaf texture features, and the plants are classified based on leaf recognition. Kayabasi utilizes the artificial bee colony optimization algorithm to train the conventional artificial neural network (ANN) to classify the wheat grains into two groups, the bread and the durum [6]. In this method, five grains' geometric parameters are chosen as input data, including length, width, area and fullness. For the same task, Sabanci *et al.* propose an ANN model based on multilayer perceptron [7], [8]. As many as 21 appearance features are selected or constructed as input data to achieve better performance. Other techniques proposed to address this task can be found in this survey [9].

Machine learning methods also play an important role in the identification of leaf diseases [10]. Zhang *et al.* construct a deep convolutional neural network architecture for maize leaf disease recognition [11]. Different convolutional neural network-based models for diseases of mango leaves and wheat leaves identification are also proposed [12]–[14]. In [14], this paper puts forward a model using convolutional neural networks. This model is capable of identifying 13 different kinds of plant diseases from the healthy leaves, and can also distinguish plant leaves out of the surrounding environment.

In this paper, we mainly study the post-harvest sorting of apples in agriculture. The key step in sorting is to classify apples into specific categories. Manual classification not

The associate editor coordinating the review of this manuscript and approving it for publication was Min Jia.

only features a large amount of work and high intensity but also is prone to fatigue, which leads to efficiency and accuracy reduction. In contrast, more and more applications are being introduced by the automatic classification of apples' categories based on image recognition. However, apples-harvesting under natural conditions is highly possible to meet various difficulties including being occluded by branches and leaves, covered by other kinds of apples, insect pests and collisions during picking. These features have a negative impact on apple recognition and classification to some extent, therefore, in the process of apple image classification, it is also necessary to consider the effect of occlusion on image classification.

With the rapid development of big data and computing power in the past few years, deep learning technology has been more and more prevalent. And in the field of image recognition, deep learning methods have achieved countless outstanding results. It shows that the use of deep convolutional neural networks can make progress in image classification methods [15]. In this work, we built a convolutional neural network structure to fulfil an apple image classification task. Considering the limited availability of apple image datasets, we collected a certain number of apple images on the web and in our lives, and marked them. we constructed a shallow CNN to prevent the occurrence of an overfitting phenomenon, but it still achieved decent results in our test set.

Due to these advantages, the proposed networks hold the promise to generalize to diverse areas. Smart visual Internet of Things arguably belongs to such areas. In recent years, the Internet of things (IoT) technology and the 5th generation mobile networks (5G) have been rapidly developed and widely applied. [16]–[18] proposed three methods respectively to realize the 5G-based green broadband communication system with simultaneous wireless information and power transfer (SWIPT) to combine wireless information transfer (WIT) and wireless power transfer (WPT), through time switching, power splitting and the division of independent frequency domain. [19] proposed a cooperative spectrum sensing model based on Dempster-Shafer Fusion. In [20], the simultaneous cooperative spectrum sensing and energy harvesting model is proposed to improve the transmission performance of the multi-channel cognitive radio (CR). The smart visual IoT, as an important part of the new generation of information technology, is the direction of continuous exploration and development of the IoT in the industrial field [21], [22]. Taking the application in this paper as an example. The identification of apple classification under the visual IoT is a real-time classifier for the obtained apple images from local area network-connected cameras, which further guides the robot to sort different apples and for statistical analysis. As described in this manuscript, the proposed machine learning-based classification method is characterized by fewer parameters to be trained, thus speeding up the training and testing procedures and holding the promise for real-time object detection, real-time visual tracking, etc. An excellent classification method without much computation and enormous datasets presents an alternative for classification related tasks in smart visual IoT.

II. RELATED WORK

Image classification has always been a fundamental and popular research topic in the field of machine vision. Traditional machine learning-based image classification methods generally require extracting features first, and then train the models according to these features. To this end, the quality of these extracted features has a huge impact on the entire classifier. With the development of big data and computing power, deep learning is predominating in image recognition tasks. Image classification based on CNN has been widely used. Deep learning requires a huge amount of image data as the training basis. The CNN model is used to automatically extract and classify features, which avoids the adverse effects of the manually extracted unsuitable features. For this reason, CNN is selected for image classification in this paper.

The traditional machine learning classification methods mainly include Naive Bayes [23], Support Vector Machine [24] and K Nearest Neighbor [25] classification. Naive Bayes is a simple classification algorithm, the basic idea of which is to calculate the probability of occurrence of each category under the conditions of this occurrence for the item to be classified. The category with the highest probability is considered to be the category to which the item belongs. Support Vector Machine is a supervised statistical learning method that minimizes empirical errors, and maximizes the interval between these sample points of different classes, resulting in a maximum interval classifier. K Nearest Neighbor is also a relatively simple classification algorithm. The k training data which is the nearest to the data to be classified is selected, and the category of the data to be classified is obtained by taking the result of the k data, with the method of taking the average number.

One of the most typical applications of deep learning in the field of image classification is CNN. The typical application is the LeNet model for handwritten digital image recognition [26], [27]. It is a relatively simple neural network structure. A convolutional layer is connected to a pooling layer, then a fully-connected layer is used, and the final classification outputs are obtained through a softmax layer. However, after that the neural network did not progress for a long time. Until the advent of distributed computing and the era of big data, neural networks began to develop rapidly.

In 2012, Alex Krizhevsky proposed a deep neural network AlexNet [28], an 8-layer convolutional neural network, which won the first place in the computer vision field competition ILSVRC 2012. The performance of AlexNet is superior on the training set compared to the traditional image classification method.

GoogleNet [29] won the championship in the ILSVRC 2014 image classification competition. No fully-connected layer was used in the GoogleNet network, which greatly reduced the number of parameters. Inception structure was

also proposed [30]. The main contribution of the inception structure is two-fold. One is to use 1×1 convolution to perform lifting and lowering in the channel dimension, and the other is to convolve and re-aggregate simultaneously on multiple sizes.

In the same year, the VGG neural network structure was proposed by the Visual Geometry Group of Oxford University [31], the structure of which became deeper, by repeatedly stacking convolutional layers with 3×3 small convolution kernels and max-pooling layers with 2×2 pooling kernels. The 16-layer deep CNN had been successfully constructed. The VGG neural network is also one of the most common CNN models that have ever been presented.

In 2015, Microsoft Research Asia proposed a deeper neural network, namely ResNet [33], which presented the idea of residual learning. By directly passing the input information to the output, the main idea of ResNet is that there are many bypasses will input directly to the latter layer. These structures are also called shortcuts. The residual idea of the new model of the network provides a new insight for the subsequent neural network research.

Although the neural network model is constantly deepening and the learning ability keeps strengthening, in the case of limited data volume, a deep convolutional neural network may perform badly due to overfitting. Hence, in this paper, we propose a kind of shallow convolutional neural network and comprehensively take advantage of some previous excellent neural networks.

III. THE PROPOSED METHOD

In this section, we describe a novel apple image classification method. Due to the lack of open apple image datasets, we collected several apple images from real life and some agricultural websites, and tagged them manually. The amount of data in our dataset is relatively limited, but the existing network models contain a large number of parameters, which makes it necessary to avoid overfitting. The classification is mainly for several common apples which lead to a small number of categories, so the deep convolutional neural network is not introduced in our experiments.

A. NETWORK ARCHITECTURE

In this work, we propose a shallow convolutional neural network structure. The specific network architecture is shown in Figure 1. The network consists of seven convolutional layers, three max-pooling layers, and two fully-connected layers. The deeper network structure such as [33] is not selected since high-risk overfitting may occur for a small number of categories. Next, we will detail these layers separately.

For the input layer, the input data size should be fixed due to the existence of the fully-connected layer. In our case, the input size is fixed to 224×224. Following the convention in the literature, images are preprocessed by scaling up to 256×256 and clipping to 224×224 from the center.

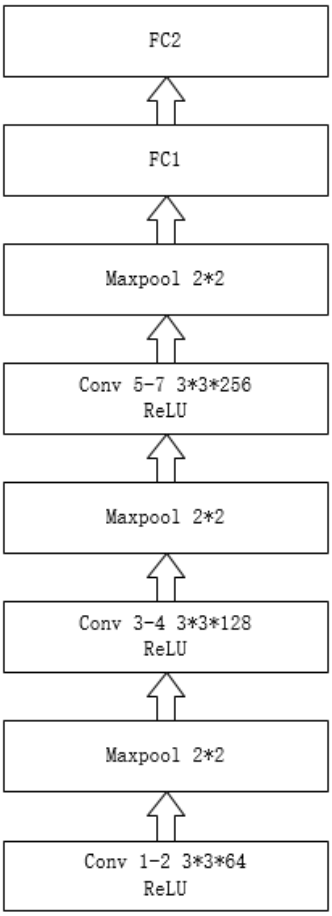


FIGURE 1. The architecture of our CNN model for classification.

Then three groups of convolution layers bond with the first input layer. The first group contains two convolutional layers with 64 filters of size 3×3, followed by a nonlinear activation function ReLU:

$$f(x) = max(0, x) \tag{1}$$

The ReLU function provides decent calculation and optimization performance with a relatively simple structure, without causing gradient vanishing and gradient explosion. Then it is followed by a max-pooling layer of size 2×2.

The second group contains two convolutional layers with 128 filters of size 3×3, also followed by a nonlinear activation function ReLU and a max-pooling layer of size 2×2.

The third group contains three convolutional layers with 256 filters of size 3×3, the structure of this group follows the previous one. In each group, several 3×3 small convolution filters of the series superposition replace the large convolution filters, which not only reduces the parameters, but also obtains two nonlinear activation and improves the expression of features. And the max-pooling layer is superimposed to highlight the obvious features of the apple images and avoid the fuzzy problem of average pooling features.

The third pooling layer is followed by two continuous fully-connected layers which improve the nonlinear expression ability of the model. The first fully-connected layer contains 1000 neurons followed by a ReLU function and a dropout layer, and the last fully-connected layer maps to the six categories of apple images.

At the end of this structure, the last fully-connected layer is followed by the softmax layer for final classification.

B. INITIALIZATION

The weight initialization method of the neural network has a crucial impact on the convergence speed and performance of the model. We use parameter random initialization. The parameters obey a Gaussian distribution with a mean of 0 and a variance of 1, which successfully initialized the parameter values to a small random number close to 0. With the parameter initialization shown in Figure 2, the output value approaches 0 quickly as the number of layers increases. When the number of layers increases, the gradient gets smaller, making the parameters difficult to update. However, this forms no negative impact on the proposed method with a small number of CNN network layers.

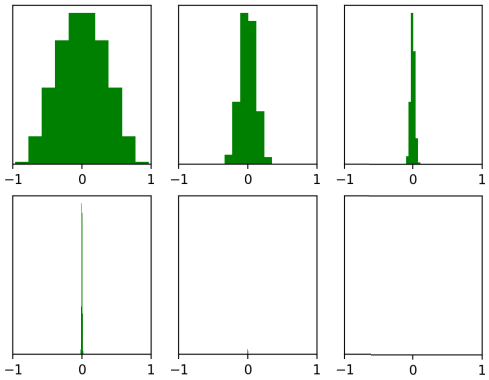


FIGURE 2. Parameter values histogram with weight random initialization.

C. TRAINING

In the training procedure, the learning rate indicates the adjustment direction of hyperparameter of the network through the gradient of the loss function. The lower the learning rate, the slower the training loss decreases. So the learning rate dynamic adjustment scheme is adopted to achieve the best effect of apple feature learning. We set the initial value of the learning rate to 0.01, and the learning rate is multiplied by 0.1 after every 10,000 iterations.

In order to prevent overfitting, a dropout layer is introduced in our network. In the forward propagation, the neural network unit is temporarily discarded from the network with a certain probability. This method can ignore some irrelevant features, reduce data volume, and improve the performance of the neural network to some extent for the limited training set.

The hyperparameters combination mentioned above is the optimal combination determined after several model adjustments and pieces of training, which ensures that the model matches the classification properties of six kinds of apples best.

D. PREDICTION

The prediction of the apple category is mainly divided into two cases. Under normal circumstances, we scale the test images to 256 × 256. Test images must be cropped into 224 × 224 as same as the input size of our network model with diverse cutting methods. For each test image, we take the five clips from the upper left, lower left, upper right, lower right, and center positions as the inputs of our model, and take the prediction results of the five images. The final classification result of this test image is obtained by taking the average on the five prediction results.

While under natural harvest condition, challenges as leaves and branches occlusion, apple surface decay or coverage by other apples would lead to wrong or miss recognition. Therefore, a block method is introduced to conduct group voting to determine the category of apple for partial occlusion, as shown in Figure 3. The images of 224 × 224 input are divided into 3 × 3, 4 × 4, 5 × 5 small blocks, and each block is predicted since the proportion of the target apple is uncertain in the case of occlusion. The category with the most votes is determined as the final category of the test image.

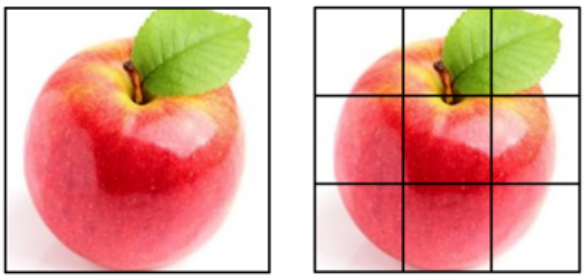


FIGURE 3. Block method for apple classification.

E. ALGORITHM COMPLEXITY

For CNN, the algorithm complexity contains two separate parts: time complexity, which is represented by floating-point operations (FLOPs), and space complexity, which is defined as the number of parameters to be trained [32]. Time complexity of a single convolutional layer is calculated as the following Equation (2).

$$Time \sim O(M^2 \cdot K^2 \cdot C_{in} \cdot C_{out}), \tag{2}$$

where M is the side length of each output feature map, K is the side length of the convolutional kernel, C_{in} is the channel number of the input, and C_{out} is the channel number of the output. The side length of feature map is determined by

the side length of input image X, kernel size K, parameter *Padding*, and stride of the kernel *Stride* according to the equation (3).

$$M = (X - K + 2 \times Padding)/Stride + 1. \tag{3}$$

The time complexity of CNN is the summary of each convolutional layer. Therefore, the time complexity of the whole CNN can be calculated as:

$$Time \sim O(\sum_{l=1}^{D} M_l^2 \cdot K_l^2 \cdot C_{l-1} \cdot C_l), \tag{4}$$

where D is the total number of layers, also known as the depth of CNN, C_l is the output channel number of the lth layer. The proposed shallow 9-layer neural network contains 7 convolutional layers and two fully-connected layers which can also be taken as a special form of the convolutional layer. After lengthy computation, the total time complexity of the proposed shallow CNN is calculated as 4.825×10^9.

The space complexity, namely, the total amount of weight parameters, can be calculated according to the following equation:

$$Space \sim O(\sum_{l=1}^{D} K_l^2 \cdot C_{l-1} \cdot C_l). \tag{5}$$

This equation is similar to the one calculating the time complexity, but discards the term M_l^2, which means the size of feature map scarcely contributes to the parameters to be trained. In this case, the proposed network arguably carries fewer parameters. After calculation, the total space complexity of the proposed shallow CNN is 1.17×10^6.

IV. EXPERIMENTS

The model is trained and tested on the Caffe framework, which is a clear and efficient deep learning framework.

A. DATASETS AND EVALUATION STANDARD

The data set mainly contains six kinds of common apples in daily life, including red Fuji, guoguang, red marshal, yellow marshal, gala and green apple, as shown in Figure 4. There are some differences among the six kinds of apples in color, shape contour and surface texture. However, they share some great similarities, as shown in Table 1. Red Fuji: large-sized, all red and form round shape. Guoguang: middle-sized, yellow mixed with green color, generally oblate shape with light red stripes on the surface. Red marshal: dark red, conical shape with a shiny surface. Yellow marshal: golden yellow, long conical shape with speckled skin. Gala: short conical with red intermittent wide stripes. Green apple: emerald green, oblate or nearly round, with dark stem and hollow. These properties constitute the particularity of the apple classification task. Visual similarities between different kinds of apples challenge our network.

The image collection mainly contains these selected apple kinds. The number of apple images of each category is about 1000. Different kinds of apples are labelled as A, B, C, D, E, and F, respectively. Due to the differences in the direction, size and brightness of the apple under natural conditions, data augmentation operations including image rotation, flip, scaling and brightness adjustment are adopted to augment the dataset. These operations not only prevent the neural network from learning irrelevant features, but also improve the diversity of training data and the adaptability of the model.

FIGURE 4. Several common apple images in daily life.

TABLE 1. Properties of six common apples.

	color	shape	surface
red Fuji	red	round	smooth
guoguang	yellow-green	oblate	light red stripe
red marshal	deep red	conical	shiny
yellow marshal	golden yellow	long conical	speckled
gala	red	short conical	wide red stripes
green apple	green	oblate	smooth

For evaluation, the confusion matrix-like table is presented, in which each row of the table indicates the proportions of images of each category that are classified into the corresponding category. Unlike confusion matrix, the probabilities are shown for the sake of clear comparison instead of the amounts of each category. In addition, based on macro-averaging, precision ratio, recall ratio and F-measure are added for more comprehensive evaluation [34]. Precision ratio P means the probability of the one considered true is really true positive. Recall ratio R indicates the ratio between predicted true samples and positive samples. F-measure F_1 is a comprehensive evaluation index, calculated as Equation (6). The performance of the model is positively correlated with the F-measure.

$$\frac{1}{F_1} = \frac{1}{2} \times (\frac{1}{P} + \frac{1}{R}) \tag{6}$$

B. EXPERIMENTAL DESIGN

ResNets are the mainstream networks in deep learning. Therefore, in addition to SVM, we also added ResNet-50 which is the classical and commonly used network structure among ResNets and ResNet-18 whose depth is close to that of the proposed shallow network to conduct comparative experiments. The architectures and complexities of ResNet-50 and ResNet-18 are described in Table 2. ResNet-50 encompasses 50 parameter layers. The input

TABLE 2. Architectures and complexities of ResNet-50, ResNet-18 and the proposed shallow CNN.

Layer name	Output size	ResNet-50	ResNet-18	Proposed
Cov1	112×112	7×7, 64, stride 2	7×7, 64, stride 2	$[3 \times 3,64] \times 2$ 2×2 max pool, stride 2
Cov2	56×56	3×3 max pool, stride 2 $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	3×3 max pool, stride 2 $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$	$[3 \times 3,128] \times 2$ 2×2 max pool, stride 2
Cov3	28×28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$	$[3 \times 3,256] \times 3$ 2×2 max pool, stride 2
Cov4	14×14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$	
Cov5	7×7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$	
	1×1	Average pool, 1000-d fc	Average pool, 1000-d fc	Average pool, 1000-d fc
	1×1	6-d fc, softmax	6-d fc, softmax	6-d fc, softmax
# params.		$\mathbf{25.5} \times 10^6$	$\mathbf{11.1} \times 10^6$	$\mathbf{1.17} \times 10^6$
FLOPs		$\mathbf{3.8} \times 10^9$	$\mathbf{1.8} \times 10^9$	$\mathbf{4.825} \times 10^9$

image is of the same size as the one in the proposed architecture: 224 × 224. The first layer contains 64 7 × 7 convolution kernels, with stride 2. So the output size is reduced to 112 × 112. Then a max-pooling layer with stride 2 is followed. After that, three consecutive building blocks are followed, with each containing three convolutional layers. These three building blocks combined are called the Conv2 layer. The Conv3, Conv4 and Conv5 consist of 4, 6 and 3 convolutional layers, respectively. The dimension increases as the layers going deeper, and reaches the peak value at the last convolutional layer. At the end of this structure, a fully-connected layer maps the network to 1000 categories, each of which corresponds to a distinct category. However, in this manuscript, only 6 kinds of apple are to be classified. So another fully-connected layer is added after this 1000-category fully-connected layer as a final classification to determine the right kind of apple. ResNet-18 is similar to ResNet-50 in structure, but with less convolutional layers and different building blocks of convolution layers.

By analyzing the complexity of the networks as shown in Table 2, the FLOPs of ResNet-50 and ResNet-18 is 3.8×10^9 and 1.8×10^9 [33], smaller than that of our shallow CNN. A question naturally arises: how come the ResNet-50 be so compact in FLOPs even with its 50 layers. It is in part due to the added two consecutive max-pooling layers prior to convolutional layers, which reduced substantially the size of the input image to a quarter of the original size. Putting two max-pooling layers at the beginning can enhance the receptive field and maintain relatively small FLOPs. However, it sacrifices the network's ability to detect detail features. Although the time complexity of the proposed network is more than two times higher than that of the ResNets, it has a great advantage in space complexity. In this case, the number of parameters, the difficulty of the training procedure and the requirements on hardware are reduced. Meanwhile, time cost of the apple recognition process is reduced, and the computational efficiency is improved as well.

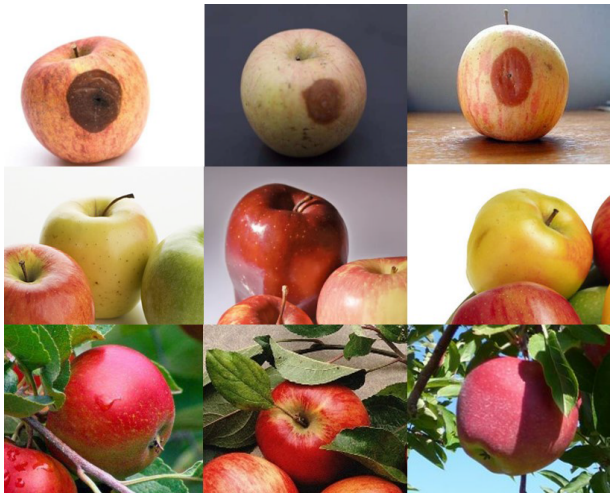


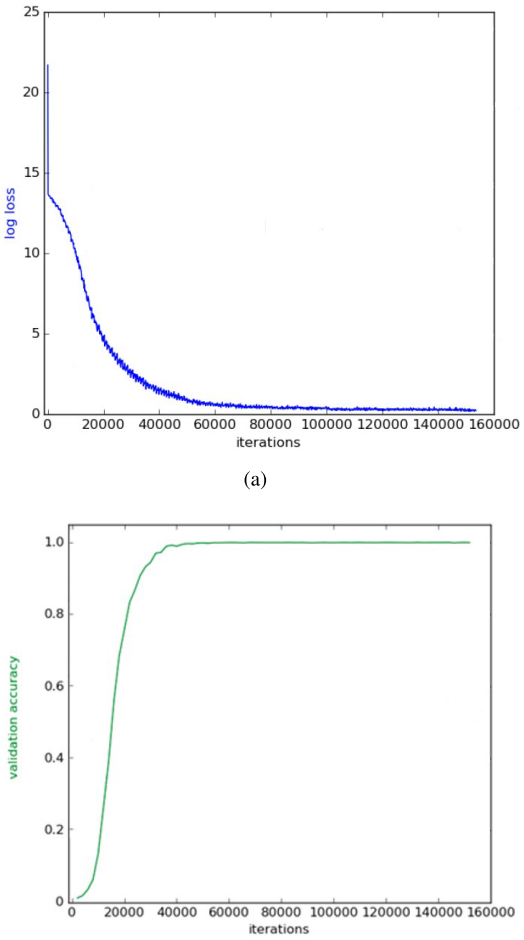
FIGURE 5. Common types of apple occlusion scenarios in the occlusion test set.

In addition, to attest the robustness of the proposed method under partial occlusion problems, another experiment is carried out to train and test the model on the occluded apple images. The images in occlusion test set are directly collected through the Internet and in real life. As shown in Figure 5, three typical types of occlusion are presented, including: 1) occlusion caused by rotten spot, 2) occlusion by other kinds of apples, and 3) occlusion by leaves and shadow. The occlusion test set encompasses over 300 images of various occlusion scenarios.

C. RESULTS

The training curves of this model are shown in Figure 6, including a training log loss curve and a validation accuracy curve. The graph proves the results that it begin to converge after about 40,000 iterations.

The confusion matrix-like tables of the four comparative experiments are shown in Table 3, Table 4, Table 5, and



(a)

(b)

FIGURE 6. (a) Log loss versus training iterations. (b) Validation accuracy in training set versus training iterations.

TABLE 3. Experiment results of SVM method.

	A	B	C	D	E	F
A	82.8%	4.6%	3.1%	3.2%	0.9%	5.4%
B	8.1%	79.6%	3.3%	2.1%	1.6%	5.3%
C	10.1%	4.6%	75.1%	5.3%	1.1%	3.8%
D	3.0%	2.7%	1.3%	85.0%	2.1%	5.9%
E	1.1%	2.3%	3.1%	3.2%	86.3%	4.0%
F	5.9%	2.1%	8.0%	2.2%	3.3%	78.5%

Table 6, respectively. Each row of the table indicates the proportions of images of each category that are classified into the corresponding category. Furthermore, more comprehensive evaluation indexes are calculated to provide more insights into the performance differences in Table 7. It presents that our proposed method achieves superior classification performance than the other three methods.

The images of occlusion test set are divided into four groups according to the occlusion ratio, and each group of test images is input into the newly trained model to obtain the classification accuracy under different occlusion ratios.

TABLE 4. Experiment results of ResNet-50.

	A	B	C	D	E	F
A	80.1%	5.0%	4.1%	5.1%	2.8%	2.9%
B	9.1%	73.2%	6.4%	5.2%	3.1%	3.0%
C	10.6%	6.8%	68.5%	4.3%	2.1%	7.7%
D	5.6%	3.4%	5.3%	80.2%	2.1%	3.4%
E	3.3%	5.3%	2.6%	6.1%	81.1%	1.6%
F	4.3%	8.1%	2.2%	5.1%	2.6%	77.7%

TABLE 5. Experiment results of ResNet-18.

	A	B	C	D	E	F
A	87.2%	3.5%	2.8%	3.9%	0.7%	1.9%
B	6.8%	85%	2.7%	2.4%	1.3%	1.8%
C	8.9%	3.6%	79.6%	2.5%	1%	4.4%
D	4.1%	1.5%	2.3%	88.5%	1.5%	2.1%
E	1.8%	2.7%	1.2%	3.1%	90.3%	0.9%
F	3.5%	5.1%	2.2%	3.2%	2.5%	83.5%

TABLE 6. Experiment results of our proposed method.

	A	B	C	D	E	F
A	90.4%	2.8%	3.1%	1.0%	0.3%	2.4%
B	2.6%	90.1%	2.5%	0.9%	0.7%	3.2%
C	4.3%	3.1%	88.5%	0.9%	0.2%	3.0%
D	1.0%	0.4%	0.6%	96.0%	1.5%	0.5%
E	0.8%	0.5%	0.6%	1.8%	95.3%	1.0%
F	2.1%	2.4%	2.9%	1.3%	0.3%	91.0%

TABLE 7. Performance comparison of SVM, ResNet-50, ResNet-18 and the proposed shallow network.

	Precision Ratio	Recall Ratio	F-measure
SVM	81.43%	81.22%	81.325%
ResNet-50	77.08%	76.80%	76.940%
ResNet-18	85.94%	85.68%	85.810%
The proposed	91.89%	91.88%	91.885%

TABLE 8. Experiment results with part occlusion.

Occlusion ratio	accuracy
0%-10%	91.8%
10%-30%	86.6%
30%-50%	72.8%
>50%	30.2%

As shown in Table 8, the classification accuracy decreases as the occlusion ratio increases. When the occlusion ratio is less than 50%, the classification accuracy can satisfy the requirements. Therefore, the proposed network and judgment scheme can effectively solve the partial occlusion problems in apple's classification task and ensure the classification accuracy.

D. DISCUSSION

Neural network's learning capacity enhances significantly with an increasing number of layers. However, if the learning capacity is strong with the limited training set, the so-called overfitting problem arises, which means the neural network is so well-trained that it performs greatly in the training set while performs badly in the test set. This phenomenon is

explained as the network overlearns the unwanted details of the training set so that its generalization performance is degraded.

The ingenious network, ResNet, is proposed to address two well-known issues challenging deep neural networks, namely, degradation problem and gradient vanishing problem. The former features a phenomenon that with the increasing layers of the deep neural network, the training accuracy declined. The latter is characterized by an extremely low update ratio of those hidden layers close to input layers, which caused the poor training results. However, when the training set is much smaller compared to the learning capacity of ResNet, overfitting problem is still inevitable. Unfortunately, the dataset we constructed is apparently under such a case from the results in this paper.

Under the background of smart visual IoT, many applications require specific tasks. Just as the apple's classification tasks, it demands the construction of its own training and test sets. Such specific tasks are not common issues we all encounter, so no more large-scale datasets are available. To construct the dataset, we collected and labelled only around six thousand apple images, and data augmentation approaches were performed to augment the dataset to about 60,000 images. Overfitting occurs with such a miniature dataset even ResNets can not avoid. It presents that the classification accuracy of ResNet-50 is about 15% lower than that of the shallow network mentioned in the paper. Overfitting decreases when ResNet-18 is applied for apple classification task, but its accuracy is still lower than the network we proposed. In addition, the residual network structure is designed to reduce degradation and gradient vanishing as the network deepens. It is mainly to ensure that when the shallow network reaches saturation accuracy, the increase of network layers will not lead to an increase of errors on the training set. However, network deepening is to obtain advanced semantic features. In our case, we pay more attention to the features of edge, color, shape and texture, rather than high-level semantic features. Therefore, the residual structure does not achieve impressive performance in our task. Under such a circumstance, the proposed shallow network has a great advantage with stronger generalization ability and higher accuracy of apple classification.

V. CONCLUSIONS

There exist miscellaneous solutions to the problem of image classification, ranging from traditional machine learning to deep learning. However, they all have their own limits. Traditional methods need to manually extract features from images. The extracted features directly determine the classification accuracy. Instead, deep learning-based methods utilize the network models to automatically extract features and perform classification, but they require a multitude of data support for training the neural network models. Nevertheless, large amounts of data may not be available for a specific task when the problem is not prevalent, e.g., the apple classification task in this paper. Manual dataset construction becomes necessary and shallow CNN deserves investigation to maintain learning capacity while avoiding overfitting due to limited data. It is from such a standpoint that we proposed this shallow network.

For apple classification task, we collected a certain number of apple images in life and on the Internet, and marked each image with corresponding category. In order to prevent overfitting problems caused by limited data volume, a shallow convolutional neural network structure is proposed. The proposed network is a network designed and adjusted on the basis of CNN architecture, but in each step, the size of the apple dataset, the properties of six kinds of apples and the problem of partial occlusion are fully considered. The advantages of each network layer are also fully utilized. Combined with the actual problems of this paper, datasets are divided into two parts, one for training and the other for testing. Through over 150,000 training iterations, the classifier with validation accuracy close to 100% is obtained.

The experimental results show that in the case of the limited apple dataset we constructed, the classification model trained by our method achieves higher classification accuracy than the traditional SVM and deep residual network. The performance achieved by deep residual network in our test set is not relatively satisfactory, which may be affected by overfitting due to too many layers but limited data. Furthermore, the residual structure does not play a very significant role in our task, which is relatively redundant. To this end, the ResNet is not an optimal choice for the apple classification task with a limited dataset we constructed. To address the partial occlusion problems in apple's classification task, we also test our model with occluded apple images using the method of block voting, and desirable results can be achieved.

The proposed shallow network is characterized by small amounts of parameters and shows resistance to overfitting with a limited dataset and has stronger generalization ability. For specific apple classification problems and based on the training and test sets that we collected and labelled, the proposed architecture performs better than SVM, ResNet-50 and ResNet-18, which verifies the effectiveness and feasibility of this shallow neural network architecture. It solves the special problems in the real apple classification task, including the interference factors caused by high similarities between different kinds of apples, the long processing time of apple recognition, and the partial occlusion problems. Such a network presents an alternative for classification related tasks in smart visual IoT and brings attention to reducing the complexity of deep neural networks while maintaining their strength.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable suggestions and queries that helped to improve this article.

REFERENCES

[1] F. Sadjadi, "Applications of image understanding technology in precision agriculture: Weed classification, and crop row guidance," in *Proc. 3rd Int. Conf. Precis. Agricult.*, 1996, pp. 779–784.

[2] P. Eddy, A. Smith, B. Hill, D. Peddle, C. Coburn, and R. Blackshaw, "Comparison of neural network and maximum likelihood high resolution image classification for weed detection in crops: Applications in precision agriculture," in *Proc. IEEE Int. Symp. Geosci. Remote Sens.*, vols. 1–8, Jul. 2006, pp. 116–119.

[3] A. J. Ishak, A. Hussain, and M. M. Mustafa, "Weed image classification using Gabor wavelet and gradient field distribution," *Comput. Electron. Agricult.*, vol. 66, no. 1, pp. 53–61, Apr. 2009.

[4] M. K. Tripathi and D. D. Maktedar, "Recent machine learning based approaches for disease detection and classification of agricultural products," in *Proc. Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2016, pp. 1–6.

[5] A. Ehsanirad, "Plant classification based on leaf recognition," *Int. J. Comput. Sci. Inf. Secur.*, vol. 8, no. 4, pp. 78–81, 2010.

[6] A. Kayabasi, "An application of ANN trained by ABC algorithm for classification of wheat grains," *Int. J. Intell. Syst. Appl. Eng.*, vol. 1, no. 6, pp. 85–91, Mar. 2018.

[7] K. Sabanci, A. Kayabasi, and A. Toktas, "Computer vision-based method for classification of wheat grains using artificial neural network," *J. Sci. Food Agricult.*, vol. 97, no. 8, pp. 2588–2593, Jun. 2017.

[8] K. Sabanci, A. Toktas, and A. Kayabasi, "Grain classifier with computer vision using adaptive neuro-fuzzy inference system," *J. Sci. Food Agricult.*, vol. 97, no. 12, pp. 3994–4000, Sep. 2017.

[9] A. Kayabasi, A. Toktas, K. Sabanci, and E. Yigit, "Automatic classification of agricultural grains: Comparison of neural networks," *Neural Netw. World*, vol. 28, no. 3, pp. 213–224, 2018.

[10] E. Yigit, K. Sabanci, A. Toktas, and A. Kayabasi, "A study on visual features of leaves in plant identification using artificial intelligence techniques," *Comput. Electron. Agricult.*, vol. 156, pp. 369–377, Jan. 2019.

[11] X. Zhang, Y. Qiao, F. Meng, C. Fan, and M. Zhang, "Identification of maize leaf diseases using improved deep convolutional neural networks," *IEEE Access*, vol. 6, pp. 30370–30377, 2018.

[12] U. P. Singh, S. S. Chouhan, S. Jain, and S. Jain, "Multilayer convolution neural network for the classification of mango leaves infected by anthracnose disease," *IEEE Access*, vol. 7, pp. 43721–43729, 2019.

[13] Z. Lin, S. Mu, F. Huang, K. A. Mateen, M. Wang, W. Gao, and J. Jia, "A unified matrix-based convolutional neural network for fine-grained image classification of wheat leaf diseases," *IEEE Access*, vol. 7, pp. 11570–11590, 2019.

[14] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep neural networks based recognition of plant diseases by leaf image classification," *Comput. Intell. Neurosci.*, vol. 2016, pp. 1–11, Jun. 2016.

[15] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and P. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580*. [Online]. Available: https://arxiv.org/abs/1207.0580

[16] X. Liu, X. Zhang, M. Jia, L. Fan, W. Lu, and X. Zhai, "5G-based green broadband communication system design with simultaneous wireless information and power transfer," *Phys. Commun.*, vol. 28, pp. 130–137, Jun. 2018.

[17] X. Liu, M. Jia, X. Zhang, and W. Lu, "A novel multichannel Internet of Things based on dynamic spectrum sharing in 5G communication," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 5962–5970, Aug. 2019.

[18] X. Liu and X. Zhang, "Rate and energy efficiency improvements for 5G-based IoT with simultaneous transfer," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 5971–5980, Aug. 2019.

[19] X. Liu, M. Jia, Z. Na, W. Lu, and F. Li, "Multi-modal cooperative spectrum sensing based on Dempster-Shafer fusion in 5G-based cognitive radio," *IEEE Access*, vol. 6, pp. 199–208, 2018.

[20] X. Liu, F. Li, and Z. Na, "Optimal resource allocation in simultaneous cooperative spectrum sensing and energy harvesting for multichannel cognitive radio," *IEEE Access*, vol. 5, pp. 3801–3812, 2017.

[21] X. Zhang, X. Wang, and Y. Jia, "The visual Internet of Things system based on depth camera," in *Proc. Chin. Intell. Automat. Conf., Intell. Automat. Intell. Technol. Syst.*, vol. 255, 2013, pp. 447–455.

[22] Q. Li, H. Cheng, Y. Zhou, and G. Huo, "Road vehicle monitoring system based on intelligent visual Internet of Things," *J. Sensors*, vol. 2015, pp. 1–16, Jul. 2015.

[23] I. Kononenko, "Semi-naive Bayesian classifier," in *Proc. Eur. Working Session Learn.*, in Lecture Notes in Computer Science, vol. 482, 1991, pp. 206–219.

[24] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.

[25] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[26] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[32] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5353–5360.

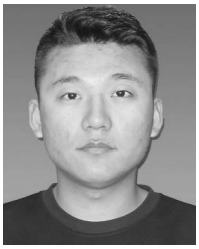
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] Z. Wang and B. Song, "Research on hot news classification algorithm based on deep learning," in *Proc. IEEE 3rd Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, Mar. 2019, pp. 2376–2380.

JINQUAN LI (Member, IEEE) received the B.E. degree in mechanical engineering from the Shaanxi University of Science and Technology, Xi'an, China, in 1992, and the Ph.D. degree in mechanical engineering from the Beijing Institute of Technology, Beijing, China, in 2004. He was a Postdoctoral Researcher with the Department of Precision Instrument and Mechanology, Tsinghua University, Beijing. He is currently an Associate Professor at the School of Automation, Beijing University of Posts and Telecommunications, Beijing. His research interests include intelligent equipment design, image processing, and FEA method.

SHANSHAN XIE received the B.E. degree in information engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2018, where she is currently pursuing the master's degree in mechanical engineering with the School of Automation. Her research interests include machine vision and target tracking.

ZHE CHEN received the B.E. degree in mechanical engineering from China Agriculture University, Beijing, China, in 2016. He is currently pursuing the master's degree in mechanical engineering with the School of Automation, Beijing University of Posts and Telecommunications, Beijing. He has been a Visiting Student at the Department of Mechanical Engineering, Tsinghua University, Beijing, since 2018. His research interests include machine vision, visual servo control, and intelligent robotics.



HONGWEN LIU received the B.E. degree in computer science and technology from Shenyang Ligong University, Shenyang, China, in 2016. He is currently pursuing the master's degree in logistics engineering with the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include automatic logistics sorting and image recognition.

JIA KANG received the B.E. degree in electronic information engineering from Northeastern University, Shenyang, China. She is currently pursuing the master's degree in mechanical engineering with the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China. She has been a Visiting Student at the Department of Mechanical Engineering, Tsinghua University, Beijing, since 2019. Her research interests include machine vision, artificial intelligence, and intelligent robotics.

ZIXUAN FAN received the B.E. degree from the Southern University of Science and Technology, Shenzhen, China. He is currently pursuing the master's degree in mechanical engineering with the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China. He has been a Visiting Student at the Department of Mechanical Engineering, Southern University of Science and Technology, Shenzhen, since 2019. His research interests include human–robot interaction and control in wearable robotics and intelligent robotics.

WENJIE LI received the B.E. degree in vehicle engineering from China Agriculture University, Beijing, China, in 2016. He is currently pursuing the master's degree in mechanical engineering with the School of Automation, Beijing University of Posts and Telecommunications, Beijing. His research interests include surface electromyograph motion recognition, embedded control, and intelligent robotics.