Detection and Recognition of Flower Image Based on SSD network in Video Stream

Mengxiao Tian¹, Hong Chen¹a, Qing Wang¹

¹China Agricultural University, 100083 No. 17 Qinghua East Road, China

Abstract. At present, most flower images could only be recognized but not detected. They can only be used in the scenes with a single target instead of the scenes with two or more targets. Some application scenarios require the human-computer interaction mode with the current location information of flowers; moreover, due to the complexity of the environment and the similarity and difference between flowers, the traditional computer visual methods are inefficient and inaccurate. Therefore, this study introduced SSD deep learning technology into the field of flower detection and identification. The flower data set published by Oxford University was used as the research object, and it was used as the input of the neural network model for training and testing. The experimental results show that the average accuracy is 83.64% based on the evaluation standard of Pascal VOC2007, and 87.4% based on the evaluation standard of Pascal VOC2012. The processing time of an image on PC is 0.13s, which indicates that high-quality automatic detection and recognition can be performed, which can facilitate the retrieval of agricultural plant information database and help people to popularize related information of flowers.

1. Introduction

With the development of artificial intelligence technology, the multimedia resources that people can obtain are no longer simple text information, but also images and video that contain more abundant information. People use pictures or video as the carrier to obtain the information of flower varieties and related attribute information through the traditional computer visual method to realize the effective breeding of flowers and help users to popularize the knowledge of the flower varieties. A large number of flower images are involved in a video, and there are multiple targets, which need to be detected and identified. The current computer visual methods can only recognize but cannot detect most flowers, and they are only applicable to the problem of single target scene, but not applicable to the actual application scene with two or more targets. Especially in the field of Internet mobile terminal, VR/AR requires a human-computer interaction mode for target detection of flowers, and flowers appearing in the APP of the image/video also need to be searched for positioning; traditional computer visual methods need complex design skills, tedious process, independent on the experience of the researchers to the characteristics of manual selection, and due to the goal of flowers and light conditions change, the diversity of background environment uncontrollable factors such as characteristics of the design of the model is not a strong generalization ability, which will affect the recognition of target, these problems cannot be solved perfectly. Due to the rapid development of deep learning in recent years, it has gradually replaced the traditional method and been applied in various fields, with obvious effects. Therefore, based on the background mentioned above, the paper studies the detection and recognition of flowers based on deep learning, and can accurately perform the real-time recognition. According to
the investigation of the agriculture of deep learning, there is no application to detect the flower variety of, only with recognition function, when there are more than one object in the image and the crisscross overlap between flowers, we need to detect the variety of flowers from the image, and get the location information of flowers in the image, attached additional identifying information on location information. The research is also an inevitable demand for deep learning and the development of smart ecological agriculture. It is a challenge in the field of agriculture to detect flower images efficiently and output identification information in real time, which has great significance and theoretical value.

In the field of target detection based on traditional methods, Navneet Dalal and Bill Triggs [1] proposed to divide the image into several regions by HOG+SVM to divide the image into several regions. The gradient edge or histogram combination of each pixel of the region species constitutes the image feature extraction of candidate regions to be manually detected. Lienhart and Maydt[2] calculated the pixel combination of each rectangle as the difference based on the rectangle adjacent to the candidate box by Harr wavelet, and then classified each block of the image. Ahonen[3] extract relevant local texture features according to image texture through the LBP algorithm. The above mentioned method requires different sliding windows of different sizes to traverse all positions due to the uncertainty of the position and size and length and width of the target in the image, it needs more time and a large number of redundant windows, which affects the detection speed and performance. Recently, the algorithm based on deep learning has become a research focus in many fields. The concept of deep learning was proposed by Hinton et al[4] in 2007, which belongs to the end-to-end real-time strategy with characteristics automatically data learning. The mainstream target detection and recognition methods based on deep learning include RCNN[5], Fast-RCNN [6], MaskRCNN[7], SSD[8] and other target detection frameworks. The target detection frameworks, RCNN and Fast-RCNN, takes up a lot of time, occupies a large amount of disk space, and cannot achieve the real-time effect due to the low speed. Every candidate region should be extracted by CNN. Mask-RCNN has extremely high requirements for equipment, and the model training is time-consuming, requiring 8 GPUs for training to achieve the exceptional effect. Different from the previous annotate method of target detection, it uses pixel-level annotation, which is very troublesome. There are no such equipment conditions and sufficient manpower in normal laboratories to execute the annotation. Compared with the above target detection framework, SSD framework relies less on feature extraction network, has better performance in large object recognition and detection, performs better in computational performance, and requires less memory.

In view of the limitations of current research methods and the complexity of image data, this paper uses SSD neural network based method for flower detection and recognition, and adjusts the parameters, as shown in Fig.1, using a 300 × 300 image as input in model, because the neural network model requires a fixed input dimension, positioning and recognition are performed simultaneously, and we make a series of adjustments.

![Fig.1 The pipeline of flower detection and recognition.](image)

2. Materials and Methods

2.1. Flower image dataset and data processing
The paper adopted the current mainstream flower data set of Oxford University, we chose 19 common flower images from the data set. Some flower images as shown in Fig. 2, as training data of deep learning.
model, there were 40 ~ 82 pieces, every kind of image reference Pascal VOC annotation and design of a data set that we annotated the target category and location on every picture, which annotated the top left corner and the bottom right corner, target information tagging of each image was annotated in the corresponding XML file, as shown in Fig. 3. Since deep learning models generally require a large number of sample data for training, to improve the robust and recognition rate of the model, it can also prevent the model from over-fitting. The amount of data collected in this paper is quite small, therefore we need to expand data sets through data enhancements, we rotated each image four angles respectively (90°, 180°, 270°, 360°) and flipped horizontal, as shown in Fig. 3, every image become eight ones, each category increased from 40 ~ 82 to 320 ~ 656, flower data set was increased from 1098 to 8284 pieces in total, 5800 of them were randomly selected as the training set, 2984 of them were randomly selected as the test set, no repeat between two sets of images. Due to expanding the operation of the data, the corresponding bounding box of expanded data sets with the original bounding box is no longer corresponding, so we need to adjust corresponding bounding box by coordinates calculation.

![Fig.2 Examples of flowers.](image-url)

![Fig.3 Data augmentation and xml annotation.](image-url)

### 2.2. SSD network structure

SSD was proposed in 2016. It uses the regression idea to obtain candidate boxes and efficiently detects objects of different sizes according to anchor mechanism. The whole training process is a one-time forward propagation process, according to the predicted detection frame and location loss of true value detection loss and recognition classification loss Back Propagation (BP), to update the weights, not all connection layer lead to target positioning using information already set test box, instead of information of the whole image, it promotes the forecast accuracy of the test box and shortens the time of target
detection, speed and performance are improved a lot compared with the previous method, making the training process and the optimization strategy convenient as can be shown in Fig.4.

Fig.4 The structure of SSD.

The input image of this network model is a 300x300 three-channel image, which is composed of VGG16[9] network structure and multiple convolution layers. The VGG16 network structure is a network used to extract image features, which is often widely used in feature extraction and has a good effect, and is widely concerned and welcomed by researchers. Replacing the full connection layer with the following convolution layer of five layers, which is faster and has fewer parameters than the full connection layer. During the training process, the model extracts the features by 3 x 3 sliding window, and then return to calculate the location of the object and category information, and its anchor working on different characteristic figure, each characteristic figure has the different feeling rages of 3 x 3 sliding window, ongoing feature extraction, then finally output multi characteristic figures of 256 channels with different scales, and various aspect ratio and size was set up in these characteristics graph, the whole network structure match each layer ReLU linear unit as the activation function. The SSD network model calculates the category probability of the target for all predicted frames, and corrects the predicted border to make it closer to the object contour. For targets of different sizes, it predicts multiple feature graphs at different scales. Due to eliminate the bounding box of alternative calculations and pixel resampling stage or characteristics, and with multiple width ratio of different categories of objects in the receptive field to calculate forecast frame and its offset, receptive field was used on these different characteristics figure later of the network, for multi-scale detection, to achieve the effect of high precision detection.

In the experiment of this paper, we finetune VGG16 as a well-trained model, and fix the underlying parameters, define different learning rates for other levels, and adjust and optimize the parameters.

2.3. Loss function
During the training process of SSD model, make each of the prior box regress to model even regress to the ground way box, loss function needs to be set in the process of debugging, it is used to calculate the error between the predicted value and true values, then guidance model learning features, the direction of the formula is as follows, made up by positioning loss and classification loss of each frame.

\[
L(x, c, l, g) = \frac{1}{N} (L_{\text{conf}}(x, c) + \alpha L_{\text{loc}}(x, l, g)) \tag{1}
\]

Where \( x \) represents the input of the model, \( c \) represents the category label of flowers, \( l \) represents the candidate box of predicted model, and \( g \) represents the true-value annotated frame of flowers; \( N \) is the number of matches; \( L_{\text{conf}} \) represents the loss of confidence, and the probability of predicted category is calculated from Softmax. \( L_{\text{loc}} \) represents the loss of position, the smoothness loss of a predicted frame and a truth frame about whether a category matches.

2.4. The evaluation criteria
The model in the paper takes Mean Average Precision (mAP) and Precision and Recall as the standards to measure the performance of the algorithm. Intersection over union (IOU) overlapping rate is generally set to be greater than 50% when we perform the detection, i.e., the ratio of intersection and union of the
detected frame and the area of the true-value frame is more than 50%, and this kind of target can be considered to be detected; otherwise, it is undetected. It is often used to calculate an criteria of the overlapping region of two target frames. Precision is the percentage of positive samples are correctly identified as positive samples in the identified images. The formula is as follows:

$$\text{precision} = \frac{Tp}{Tp + Fp} = \frac{Tp}{n}$$  

(2)

Recall is the ratio of all positive samples in the test set that can be correctly recognized as positive samples. The formula is as follows:

$$\text{recall} = \frac{Tp}{Tp + Fn}$$  

(3)

Tp in formula 2 and formula 3 is true positives, positive samples are correctly identified as positive samples, and the number of flower categories is detected correctly; Tn is true negatives, negative samples are correctly identify negative samples, the number of judged as the background; Fp is false positives, negative samples are wrongly identified as positive samples, and the number of detected flowers categories is mistaken for other flowers; Fn is the false negatives and false negative samples, the positive samples is error recognition as negative samples, the number of undetected flower category. When the Recall value keeps increasing, the Precision value stays at a high level, so the performance of the classifier is getting better. MAP is the average of multiple categories of AP, which is the area below the Precision-recall curve. Generally speaking, the better a classifier is, the higher the AP value is. The size range of mAP is in the interval of $[0, 1]$, and the larger the value is, the better the model performance will be. Taking Pascal voc data set as the evaluation standard, average precision AP is the average value of recall rate $r \in [0, 1]$ precision $p(r)$. Interpolation method is used to calculate $P_{\text{interp}}(r)$. For a certain recall($r$), the precision value takes out the maximum precision of all recall $R$ to ensure that the precision-recall trend is monotonically decreasing.

$$\text{AP} = \frac{1}{11} \sum_{r=0.0, 0.1,...,1} P_{\text{interp}}(r)$$  

(4)

$$P_{\text{interp}}(r) = \max p(r)$$  

(5)

3. Experiment set up

In this paper, we built a convolutional neural network model based on the TensorFlow deep learning framework, and trained the model by loading the pre-trained VGG16 model. The Adam optimization function is used as the gradient descent algorithm of the model to make the model converge faster, and the weight initial test rate is set to 0.001, and the weighting attenuation coefficient of regularization is set to 0.0005, and the attenuation factor of learning rate is set to 0.94. The NVIDIA GeForce GTX 1060 is used as the working platform, the operating system is Ubuntu 16.04, and the memory is 8G.

4. Experimental results and analysis

4.1. Curve of Training Loss

Fig. 5 shows the loss curve for training, where the abscissa represents the number of steps in the run and the ordinate represents the loss of training, which is that the error between the true and predicted values. It can be seen that the network is rapidly converging; the model of this paper has received the training of 35000 steps, the convergence effect of the model is relatively stable, and the deviation of the initialization can be repaired. Convolutional neural networks can automatically learn image features without the need to manually extract features.
4.2. Overall comparison

Through a series of adjustment operations, the trained model is validated on the test set. Whether the scheme model is effective and robust, the experiment analyzes the results of each type of flower and achieves the target detection. The average accuracy values for a category are shown in Table 1:

| Category                              | Number | VOC2007(AP) | VOC2012(AP) |
|---------------------------------------|--------|-------------|-------------|
| Balloon flower                        | 119    | 79.82%      | 83.17%      |
| Pink primrose                         | 102    | 86.23%      | 90.13%      |
| Hard-leaved pocket orchid             | 159    | 95.2%       | 99.38%      |
| Canterbury bells                      | 107    | 34.56%      | 31.73%      |
| Sweet pea                            | 164    | 51.24%      | 50.91%      |
| English marigold                      | 193    | 95.0%       | 99.0%       |
| Tiger lily                            | 111    | 90.91%      | 96.99%      |
| Moon orchid                           | 104    | 88.84%      | 90.96%      |
| Bird of paradise                      | 240    | 90.91%      | 99.55%      |
| Monkshood                             | 128    | 58.33%      | 59.96%      |
| Globe thistle                         | 115    | 96.41%      | 99.26%      |
| Snapdragon                            | 232    | 82.6%       | 87.37%      |
| Colts foot                            | 234    | 92.88%      | 97.72%      |
| King protea                           | 130    | 88.15%      | 93.23%      |
| Spear thistle                         | 141    | 90.1%       | 96.72%      |
| Yellow iris                           | 149    | 88.41%      | 92.79%      |
| Globe-flower                          | 106    | 90.36%      | 93.75%      |
| Purple coneflower                     | 238    | 96.49%      | 99.58%      |
| Peruvian lily                          | 218    | 92.77%      | 98.76%      |

From the target test results in Table 1, except that the average accuracy of Canterbury bells, Sweet pea and Monkshood is less than 80%, the average accuracy of target detection of other flowers is as high as 80%, and some even reach 99%, the overall average accuracy is higher, it can be shown that SSD is suitable for the detection and recognition of flowers, the recognition effect is better, and it has better robustness. It can be seen from Table 2 that the speed is faster and can achieve end-to-end real-time target detection.

| dataset       | VOC2007(mAP) | VOC2012(mAP) | Cost time |
|---------------|--------------|--------------|-----------|
| Test(2990)    | 83.64%       | 87.42%       | 0.13s/an image |

4.3. Video testing
Using the trained SSD model as the recognition engine of the video stream, the GPU is used to complete the calculation process. We read the video file and then extract the key frames, which can reduce the redundant information of the video data stream and feed the training. A good model performs target detection, and each frame performs target detection, and the detected border and the identified type are additionally displayed in each frame of video.

Table 3 Examples of tests

| Bird of paradise: 0.549 | Sweet pea: 0.409 |
|------------------------|------------------|
| ![Bird of paradise](image1) | ![Sweet pea](image2) |

| Pink primrose: 0.752 | King protea: 0.725 |
|----------------------|---------------------|
| ![Pink primrose](image3) | ![King protea](image4) |

Table 3 shows the results of some representative video stream-based samples. It can be found that the model has good robustness. The value is higher, which can meet the needs of practical applications.

5. Conclusion

The research establishes the flower images detection and recognition based on SSD in video stream. The main conclusions are as follows:

1. Because most flower images can only be recognized but not detected. They can only be used in the scenes with a single target instead of the scenes with two or more targets and the traditional computer visual methods are inefficient and inaccurate. SSD deep learning technology is used in the paper to solve the problem of low efficiency and low accuracy to some extent, and the scope of application scenarios is improved at the same time.

2. The solution in the paper fully balances the accuracy and speed of the detection. It can meet the real-time requirements in the detection speed, with the small memory, so that it can also detect and identify in the low configuration. As long as there is a notebook with a video card with more than 4G memory, the detection model can be trained, it is a simple and cheap laboratory.

Acknowledgments
This work was supported by National Natural Science Foundation of China [Grant No.41601491]; The Fundamental Research Funds for the Central Universities [Grant No. 2018QC080, 2018TY002]].

References

[1] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. 2005.

[2] Lienhart R, Maydt J. An Extended Set of Haar-like Features for Rapid Object Detection[C]// Image Processing. 2002. Proceedings. 2002 International Conference on. IEEE, 2002.

[3] Ahonen T, Hadid A, Pietikainen M. Face Description with Local Binary Patterns: Application to Face Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12):2037-2041.

[4] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786):504.

[5] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2014:580-587.

[6] Girshick R. Fast R-CNN[J]. Computer Science, 2015.

[7] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[J]. 2017.

[8] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. 2015:21-37.

[9] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.