Fruit and Vegetables Classification System Using Image Saliency and Convolutional Neural Network

Guoxiang Zeng
College of Information Engineering, Communication University of China
Beijing, China
cimoaa@163.com

***Abstract*—Fruit and vegetables classification and recognition are still challenging in daily production and life. In this paper, we propose an efficient fruit and vegetables classification system using image saliency to draw the object regions and convolutional neural network (CNN) model to extract image features and implement classification. Image saliency is utilized to select main saliency regions according to saliency map. A VGG model is chosen to train for fruit and vegetables classification. Another contribution in this paper is that we establish a fruit and vegetables images database spanning 26 categories, which covers the major types in real life. Experiments are conducted on our own database, and the results show that our classification system achieves an excellent accuracy rate of 95.6%.**

Keywords-fruit and vegetables classification; image saliency; convolution neural network (CNN); VGG

I. INTRODUCTION

Fruit and vegetables classification remains challenging in image recognition in computer vision, given that fruit and vegetables may have similar colors, shapes and texture. In production and sales, it takes too much time to label or count the money for various fruit and vegetables, which works little for process automation. Besides, people pursue higher quality and efficiency of life, thus various highly automatic and efficient methods are raised for well-being. Therefore, we devise a high accuracy classification system aimed at fruit and vegetables without consuming much manual labour or time.

In recent years, researches have been devoted to fruit recognition and classification and significant progress has been made in this field. Hetal et al. [1] proposed an improved method based on multi-features for detecting the orientation of the fruit object in an input image. Woo et al. [2] devised nearest neighbors classification based on color, shape and size features. They compared test images with stored image categories and gain the classification results. Pragati et al. [3] also chose K-Nearest Neighbors (KNN) as classification component, but a new texture feature was taken into account. Zhang et al. [4] proposed a method based on fitness-scaled chaotic artificial bee colony algorithm and feedforward neural network (FSCABC-FNN). The network extracted color, shape and texture as features. The method reached a rather good accuracy. In the meanwhile, Zhang et al. [5] proposed a method based on multi-class KSVM. Color, shape and texture were extracted as features. Recently, Lukas et al. [6] explored a new method based on random forest to classify food or nonfood and food categories. They utilized random forest to mine discriminative components. They had a large food and nonfood image database, including 101 food categories. As we can see, a majority of methods on fruit classification are traditional and old-fashion. Linear classifier and KNN classifier take great part in fruit classification and features are rare, which limits the development and accuracy of methods.

In the past years, a rapid improvement and development have been gained in the field of artificial neural network. The outstanding classification accuracy and operational precision make it an alternative choice when we handle image processing problems. Neural network has numerous characters like unsupervised learning or rich extraction of features, which fairly improves the performance of network. And the convolutional neural network (CNN) makes a big success in large image process. AlexNet [7], VGG [8], GoogLeNet [9] and ResNet [10] are excellent CNN models in the past few years. These models achieve impressive effect in image process, such as image recognition and classification tasks, object detection and semantic segmentation. Alex Krizhevsky et al. [7] put forward a deep convolution network architecture in 2012, achieving outstanding performance, which re-caused a boom of CNN. Later, plenty of high quality CNN models were proposed and promoted in the field of computer vision. The winners of ILSVRC2014 went to VGG and GoogLeNet, which go deeper in their convolution architectures respectively and decrease the top-5 error again. With the deepening of architecture, the CNN model can approximate the objective function in nonlinear growth and gain a better character representation. Present deep learning models can reach a knockdown top-5 error, which is even lower than the identification error rate of the human eye. That is to say, the recognition and classification performance of presenting deep learning model is beyond human eye. What is more, CNN is applied to fruit and vegetables classification as well, which takes a fairly great effect. For instance, Ashutosh et al. [11] experimented on food or nonfood classification and food recognition utilizing a pre-trained GoogLeNet model, which showed a high accuracy. On the other hand, they established their own two databases.

As early as in 1998, the concept of saliency was put forward for the first time by Christof Koch et al. [12] Combined with an influential biologically-plausible architecture of their previous study, they proposed a model of

978-1-5090-5363-6/17/$31.00 ©2017 IEEE

saliency-based visual attention and pioneered the first. Afterwards in 2006, Jonathan Harel et al. [13] devised a saliency detection based on the image with matlab in Koch Lab. The new buttom-up visual saliency model achieved a high level

Fig. 1. Examples of images with image saliency regions with T =0.37

performance. After this, various algorithms, such as Spectral Residual [14], Phase Spectrum [15] and group saliency [16], were proposed and confirmed with excellent saliency object detection and segmentation. Image saliency intuitively depicts some part of the scene, which could be an object or area, and these parts seem to be relatively highlighted than adjacent area. To be more specific, image saliency regions make the measuring object more intuitive and clear, which contributes a lot to subsequent processing on object detection or image recognition. Fig.1 shows some preprocessed images.

Therefore, we carry out a novel fruit and vegetables classification system using image saliency and CNN model in this paper. A VGG model is chosen to classify the objects. Our VGG model is developed in convolution layers and max-pooling layers in class D, three Fully-connected (FC) layers and a soft-max layer. [9] Besides, image saliency works for the input images with the outstanding main regions. In addition, we establish our own fruit and vegetables images database as well, including 26 categories and hundreds images per category with high diversity.

The rest of the paper is organized as follows: Section 2 introduces our work in image database, the use of image saliency and VGG model. Section 3 presents the experiments and our results step by step. Section 4 draws conclusions and future tasks.

II. DATABASE, IMAGE SALIENCY AND VGG MODEL

A. *Establish database*

Fruit and vegetables recognition and classification are not easy tasks, not only because of similar fruit and vegetables categories but also owing to the lack of well-formed and available databases. The number of training samples is small in some former studies, which is not conducive to the extensibility of an algorithm. While there are well-formed databases spanning vast categories, they are private and unavailable.

Hence, we decide to establish our own fruit and vegetables images database. The fruit and vegetables database is obtained after a period of collection and filtering. First of all, we create a preliminary draft of 13 fruits and 13 vegetables. They are commonly consumed by people in markets, because our fruit and vegetables classification system aims at automating the process of labeling and counting for production and life, not for recognizing infrequent fruit and vegetables species.

The first step is online collecting. We download images from picture websites with the help of web crawler technology. On purpose, the download images will pass artificial selection to keep the database clear. Here are several principles for artificial selection: (a) There are single instance and multiple samples of objects from various images. (b) The diverse shapes and colors caused by nature or the freshness of fruit and vegetables will be tolerant. (c) We ensure there are images with objects in multiple scenes or surroundings. (d) The proportion of objects in an image will be taken into account. That is, images with too big or too small objects will be passed. (e) A vital point will be reached that half-baked fruits or vegetables will be classified to the correct category, considering that there are dissected fruit and vegetables for sale in real life. In other words, we will collect images with exposed pulp of fruit and vegetables.

Fig. 2. Examples of shooting photos

To enrich our database, the following step is shooting photos via digital camera. We take photos of the draft fruit and vegetables, considering the principles mentioned above. Moreover, the background of shooting photos is an important factor as well, because in contrast to online images the shot environment can be relatively fixed and the objects of fruit and vegetables should be more accurate. Hence, the first option is the pure white, black or gray background. The second kind of

shooting images is in supermarkets background. Furthermore, the objects in images should be highlighted. That is to say, we should ensure the objects can express the exact information of true fruit and vegetables. Some photo images are shown in fig.2.

B. Image saliency

As for our experiment object, fruit and vegetables, image saliency predicts the intuitive notice of the object of human eye. That is to say, the prediction utilizing image saliency helps us to extract image contents or areas of dense features in certain ways and filter the complicated backgrounds that may cause unnecessary noise.

We choose a buttom-up graph-based visual saliency (GBVS) model [13] to determine the significant area in an input image. A buttom-up mode is fast and related to the mechanism of visual coding. And this mode only uses local information to determine the saliency area. Three modules can quickly explain our process in extracting saliency:

- Particular features Extraction: this model chooses DKL colors, intensity and orientation as features to extract.
- Activation: activation maps are formed utilizing equilibrium distribution over map locations.
- Normalization/ Combination: the output is a single image combined with normalized activation maps.

However, the key is the GBVS saliency value matrix. Saliency values are gained via equilibrium distribution over map locations in a Markov chains over various maps. We assume that every pixel has a saliency value representing its saliency in a whole image, and the fact proved correct. We need to ensure that it is the object in image that works for classify, not the noise or interference.

Definition 1.The saliency in an image with proportion β determines which pixels will be regarded as useless for image process.

$$\beta = \frac{\sum_{(x,y)} M(s < st)}{\sum_{(x,y)} M(s)} \tag{1}$$

Definition 2. The certain saliency value st is determined by

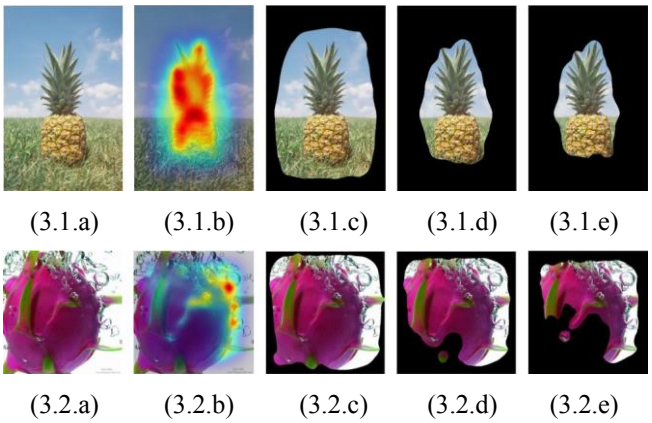
$$st = \begin{cases} \mathrm{st} + \Delta, & \beta < T \\ st, & \beta \geq T \end{cases} \tag{2}$$

Where (x, y) is a pixel in an input image, $M(s < st)$ denotes a pixel whose saliency value under the certain saliency value st, $M(s)$ shows the saliency value of each pixel, β denotes the percentage of the pixels of the whole image and Δ is a tiny increment. We calculate the amount of $M(s < st)$.When the amount takes the percentage β of the whole image, the amount can be regarded as useless and discarded. β is the key percentage to select our saliency area. The certain saliency value st is calculated during an iterative process, nor a constant value. The priori knowledge is that image saliency value of each pixel $M(s)$ is limited to $[0,1]$. So, we can set st to 0 and Δ to a tiny constant value as 0.001. The threshold T in this experiment is set to 0.37 via a large number of contrast and calculation. T can be adapted to different scenes. Sometimes when the backgrounds are complicated or the object is too big to occupy the whole image, the preprocessed image can amplify or decrease the saliency area. Some improper output images can be seen in fig.3.

C. VGG Model

VGG model is a typical CNN with high classification and recognition rate. It goes deeper than former traditional CNN architecture. To be specific, VGG model increases the depth of the network steadily by adding more convolutional layers. And very small convolution filters (3*3) make it work successfully.

The input image to our model is firstly down-sampled to a fixed-size 224*224 RGB image, and then subtracts the mean RGB value for each pixel. After that the image goes through a series of convolutional layers and max-pooling layers right after convolutional layers, where the receptive field is small (3*3) and the convolution stride is set 1. Max-pooling layers have a 2*2 window with stride 2. Two following FC layers have 4096 channels each, and the last FC layer performs 26 ways classification. The final layer is a soft-max layer. We define 26 categories fruit and vegetables and decide that our network output top-3 classification results. Though the top-1 classification result is what we want, we believe that top-3 results can fully embody the performance of our network.



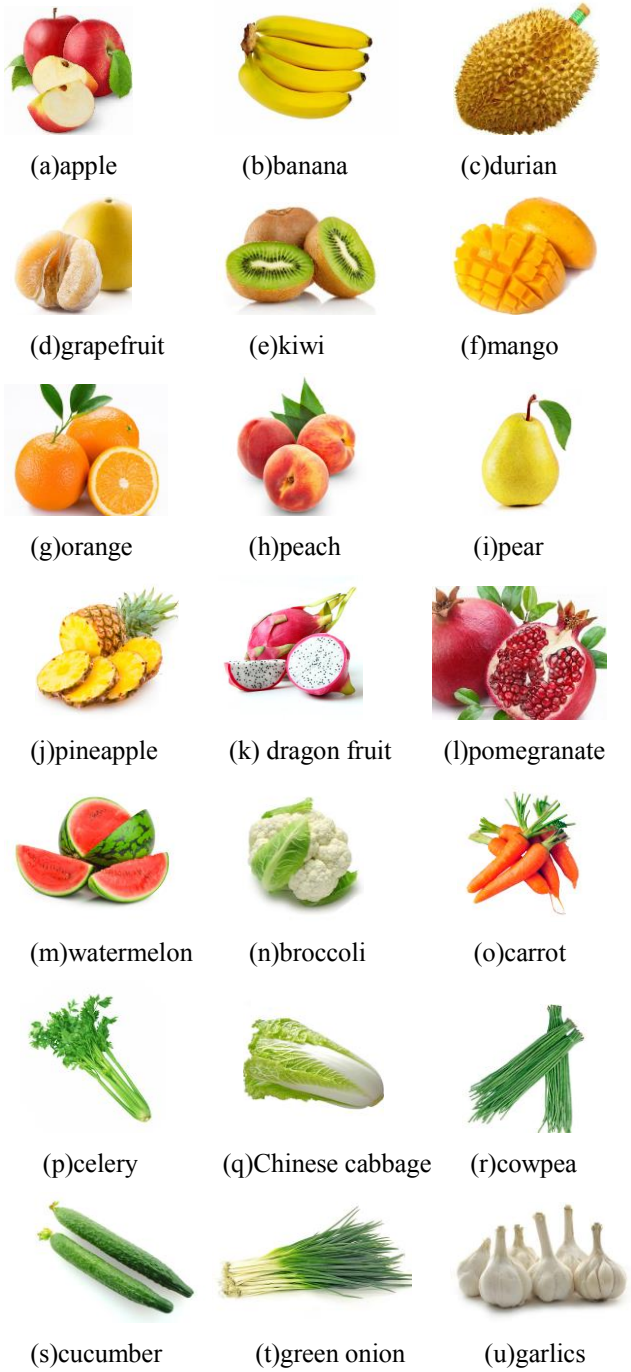
(3.1.a) (3.1.b) (3.1.c) (3.1.d) (3.1.e)

(3.2.a) (3.2.b) (3.2.c) (3.2.d) (3.2.e)

Fig. 3. Improper output images. (3.1.a) is the original image;(3.1.b) adds the saliency map; (3.1.c) is the output image through the preprocessing with T = 0.37; (3.1.d) with T = 0.56; (3.1.e) with T = 0.75; (3.2.a) is the original image;(3.2.b) adds the saliency map; (3.2.c) with T = 0.10; (3.2.d) with T = 0.20; (3.2.e) with T = 0.37.

III. EXPERIMENTS AND RESULTS

Our database consists of 12,173 images grouped into 26 categories, which covers the major types that can be commonly consumed in daily life. We list the 26 different categories here: apple(719), banana(457), durian(485), grapefruit(445), kiwi(996), mango(867), orange(523), peach(614), pear(696), pineapple(419), dragon fruit(818), pomegranate(524), watermelon(932), broccoli(278), carrot(202), celery(319), Chinese cabbage(167), cowpea(178), cucumber(328), green onion(121), garlics(285), mushroom(334), onion(347), pepper(371), pumpkin(242), tomato(506). The whole categories and samples are shown in fig.4.



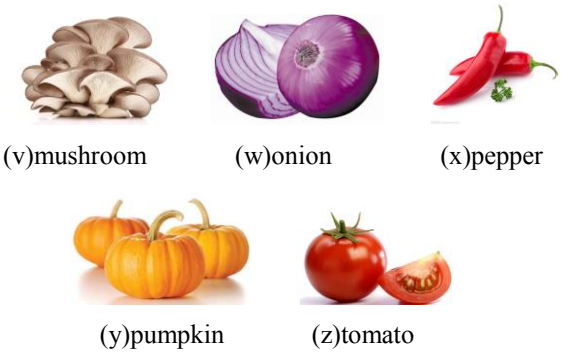


Fig. 4. Examples of fruit and vegetable database of 26 categories

For our performing experiments, images in our database are randomly divided into 80% training and 20% validation for objective and fair. We choose to train our system on GPUs. In our training process of VGG model, the batch size is set to 64. We use an initial learning rate of 0.001, and then decreased by a factor of 0.1 per 20 epochs. The number of training epochs is opt to 100.

A. Image preprocessing based on image saliency

In our study, an image preprocessing based on image saliency is added to reduce interference and noise caused by complex background on feature extraction. We design a control experiment testing the performing of our image preprocessing. The input images consist of online images and shooting photos. The VGG model with raw input images reaches a top-1 classification accuracy of 89.6%. Meanwhile, input images that are picked out the saliency regions achieve the accuracy of 92.5%. We apply the mentioned methods to our database. Compared with former methods, our system outperforms KNN by 7.2%, and is higher than KSVM by 6.3%, than FSCABC-FNN by 2.6%. Table 1 lists the accuracy of the methods. Our proposed classification system has gained a considerable recognition and classification accuracy.

B. Online collecting images and shooting photos

The advantages of online images appear in rich scenes and different forms of fruit and vegetables. A VGG model with input images consisting of online images has strong tolerance, which ensures fruit and vegetables in complex circumstances can be classified correctly. On the other hand, shooting photos via digital camera makes the concept of fruit and vegetables more distinct. We take different numbers and various forms of a fruit or vegetable category into account, which guarantees the richness of our database.

How the different source of images contributes to our classification system is what we are about to figure out in this experiment. We divide input images into two groups: (a) pure online images, (b) pure shooting photos. We decide to train the VGG model with group (a) and group (b) respectively, name the networks as VGG-O and VGG-S, and record the accuracies. After that, we apply group (b) to the pre-trained VGG-O model called VGG-OS and we then input group (a) to the pre-trained VGG-S model called VGG-SO in turn. The two experiments

are put forward to make certain whether there is a better way to train our system to gain the best accuracy. As we can see, the VGG-OS model reaches the best top-1 accuracy. To be more specific, our system can recognize and classify the correct category of fruit and vegetables with high accuracy when the network is pre-trained by images with complex surroundings and trained by images with more exact details. Table II lists TOP-1 classification accuracy in different training types to VGG model.

TABLE I. COMPARSION OF ACCURACY PERFORMANCES ON METHODS

Methods	Top-1 accuracy (%)
KNN[a]	85.3
KSVM	86.2
ABC–FNN	86.7
FSCABC–FNN	89.9
PROPOSED(R)[b]	89.6
PROPOSED(P)[c]	92.5

a. Method proposed by Woo Chao Seng et al.

b. VGG model with raw images

c. VGG model with input images picked out saliency regions

TABLE II. COMPARSION OF ACCURACY PERFORMANCES ON DIFFERENT TRAINGING IMAGES TYPES TO VGG MODEL

VGG model with different input images	Top-1 accuracy (%)
VGG-O	89.6
VGG-S	94.0
VGG-OS	95.6
VGG-SO	90.5

IV. CONCLUSION

In this paper, we propose a novel fruit and vegetables classification system based on image saliency and VGG model. On the one hand, we establish our own experiments database. On the other hand, we bring a new idea about image saliency in the field of image process. We also utilize VGG as CNN model to classify the objects. The experimental results show that our overall top-1 accuracy of 95.6% in fruit and vegetables recognition and classification outperforms vast methods. As a future direction, we aim at recognizing and classifying major fruit and vegetables consumed in daily life. We hope our method and database will convey more than this paper and inspire the future studies.

ACKNOWLEDGMENT

Zeng thanks Da Pan, Mingliang Han, Ying Xu and Zefeng Ying for their valuable suggestion and time for collecting database.

REFERENCES

[1] Hetal N.Patel, R. K. Jain and Manjunath V. Joshi. “Fruit Detection using Improved Multiple Features based Algorithm.” *International Journal of Computer Applications (2011)*, vol. 13- No. 2, pp. 1-5, January 2011.

[2] Woo Chaw Seng, and Seyed Hadi Mirisaee. “A new method for fruits recognition system.” *international conference on electrical engineering and informatics(2009)* ,pp. 130-134, September 2009.

[3] Pragati Ninawe and Mrs. Shikha Pandey. “A Completion on Fruit Recognition System Using K-Nearest Neighbors Algorithm” International Journal of Advanced Research in Computer Engineering & Technology (2014), vol. 3, Issue. 7, July 2014.

[4] Yudong Zhang, Shuihua Wang, Genlin Ji and Preetha Phillips. “Fruit classification using computer vision and feedforward neural network.” *Journal of Food Engineering (2014)*, pp. 167-177, 2014.

[5] Yudong Zhang and Lenan Wu. “Classification of Fruits Using Computer Vision and a Multiclass Support Vector Machine.” *Sensors* 12.9, pp. 12489-12505, 2012.

[6] Lukas Bossard, Matthieu Guillaumin and Luc Van Gool. “Food-101 – Mining Discriminative Components with Random Forests.” *european conference on computer vision* (2014), pp. 446-461.

[7] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. “ImageNet classification with deep convolutional neural networks.” *neural information processing systems* (2012),pp. 1097-1105.

[8] Karen Simonyan and Andrew Zisserman, “VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION” ICLR 2015.

[9] Christian Szegedy et al. "Going deeper with convolutions." *computer vision and pattern recognition* (2015), pp. 1-9, 2015.

[10] Kaiming He et al. "Deep Residual Learning for Image Recognition." *computer vision and pattern recognition* (2015), pp. 770-778, 2015.

[11] Ashutosh Singla, Lin Yuan, Touradj Ebrahimi “Food/Non-food Image Classification and Food Categorization using Pre-Trained GoogLeNet Model”

[12] Laurent Itti, Christof Koch and Ernst Niebur. “A model of saliency-based visual attention for rapid scene analysis.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.11 (1998), pp. 1254-1259, 1998.

[13] Jonathan Harel, Christof Koch and Pietro Perona. “Graph-Based Visual Saliency.” *neural information processing systems* (2006),pp. 545-552, 2006.

[14] Xiaodi Hou and Liqing Zhang. "Saliency Detection: A Spectral Residual Approach." *computer vision and pattern recognition* (2007), pp. 1-8, 2007.

[15] Chenlei Guo, Qi Ma, and Liming Zhang. "Spatio-temporal Saliency detection using phase spectrum of quaternion fourier transform." *computer vision and pattern recognition* (2008), pp. 1-8, 2008.

[16] Mingming Cheng et al. "SalientShape: group saliency in image collections." *The Visual Computer* 30.4 (2014), pp. 443-453, 2014.

[17] Hao Zhu, Biao Han, and Xiang Ruan. "Visual saliency: A manifold way of perception." *international conference on pattern recognition* (2012), pp. 2606-2609, 2012.

[18] Tsung-Yi Lin, Piotr Dollár 1, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie “Feature Pyramid Networks for Object Detection”, airXiv:1612.03144v1, [cs.CV], 9 December 2016

[19] Ruaa Adeeb Abdulmunem Al-falluji, “Color, Shape and Texture based Fruit Recognition System” International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol.5, Isuue.7, July 2016.