Study of Key Algorithm on Automatic Classification of Insect Images

LI Jian[1, a], ZHANG Lei[1, b] and YAN BaoPing[1, c]

[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China

[a]email: lijian@cnic.cn, [b]email: zhanglei@cnic.cn, [c]email: ybp@cnic.cn

Keywords: Image Similarity, Image Feature analysis, Color Histogram, Textural Features, Insect Species Identification

Abstract. Insects, due to their diversity, are the largest group of creature on earth, many of which are yet to be discovered right now. As the photographic hardware develops and gets pervasive rapidly, insect images or pictures have become an important approach for insect researchers to conduct scientific study. However, considering the huge volume of insect image data, researchers usually are not interested in all the images and instead, hope to categorize these images preliminarily and then hand them on for manual processing, which will significantly improve work efficiency and reduce manual workload. In this paper, color histogram and textural features are adopted to collect image features which then use clustering analysis method for automatic classification of insect species. The experimental result is quite good. And on the basis of this, more image features can be extracted to gain more accurate results, or program efficiency can be optimized through parallelization, during which a species feature database is gradually established to have automatic comparison and identification of insect features.

Introduction

Due to their variety in types and shapes, insects are the largest group of creature on earth, accounting for more than 50% of the total species. Many of them are still yet to be discovered currently. Whether in individual number, biomass, species or gene number, insects have a critical position in biodiversity. They have complex and close relations with human. Only by accurately identifying their exact species, can the agricultural loss from diseases and pests be reduced or rare insect species be protected and biodiversity be studied. Insect species identification is a complex and arduous task. For a long time, the work of insect classification and identification has been undertaken only by a few insect researchers and plant protection scientific staff, who mainly depend on manual examination, visual study, intuition and group experience. So, their results are usually affected by the subjective judgment of identifiers, which tends to have unstable rate of identification, in particular under the conditions of long work time and large workload and high misjudgment rate. Furthermore, considering the variety of insect types, it is hard to guarantee that identifiers can accurately identify all the species due to personal knowledge constraint. Also, the research team of insect taxonomy globally is shrinking in recent years, thus the actual needs for insect species identification are far beyond the current workload by taxonomists. These problems, to a large degree, limit the popularity of insect knowledge among the public and may cause huge loss to agricultural production and economic activity, which in turns will not help protect rare and precious insect species as well.

Along with the advancement of digital image processing technology, it has gradually played a role in insect classification and identification. Currently, a lot of research has been carried out, such as those in References [1, 2 and 3]. However, relevant research usually limits to certain insect group or scope, which, in fact is added with some man-made definition and priori knowledge. With the pervasion of photographic devices, massive insect pictures are collected and piled before insect researchers. Considering the huge volume of images, manual classification processing has got increasingly unbearable or even seriously hindered the insect research progress and discovery of new species. Therefore, this paper tries to analyze and process large volumes of insect image data by using image feature extraction and data mining algorithm so as to work on lightened workload for insect researchers.

The general technical architecture adopted in this paper can be shown in Fig. 1, which is divided into four core models, i.e. data pre-processing, image feature extraction, image feature comparison and automatic image clustering, which will be elaborated separately.

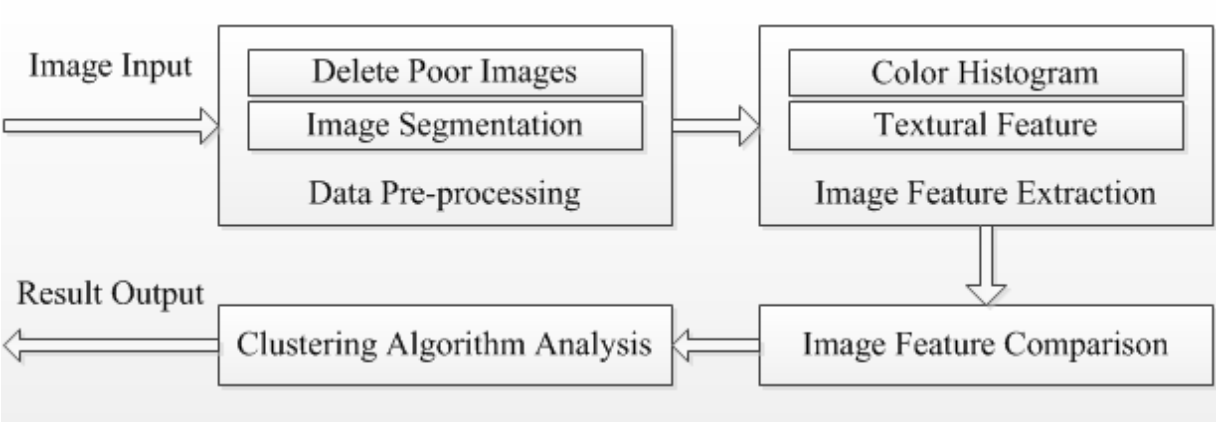

Fig. 1. General Processing Diagram

Data Pre-processing

The fast penetration of photographic devices enable a growing number of people to get involved in the activity of taking insect pictures, which come from nature, labs and samples. As the image quality varies, so it is very necessary to have pre-processing on data to ensure data quality and delete noise information (such as image background).

Delete Poor Images

Because image data comes from various channels, like Internet, insect researchers and amateur photographers, difference in photographing skills will inevitably lead to poor image quality like excessive exposure, vagueness or distortion. If these kinds of images are not deleted, they will exert negative impact on our analysis result.

Image Segmentation

An insect image contains lots of information. Besides insects, it usually covers basic data like living environment. And our automatic classification on certain species mainly depends on the insect part in the image, i.e. Region of Interest (Short for ROI). Although other parts also have the meaning as reference, it will probably create error for analysis results, so the background region needs to be deleted, which will involve the key technology of image segmentation.

Image segmentation is a process of dividing the digital image into non-intersecting (non-overlapping) regions through three different principles: region-based segmentation, which classify pixels into each object or region; boundary-based segmentation, which only needs to identify the boundary between areas; edge-based segmentation, which identifies edge pixel first and then connect them to form a boundary. Since 1970s, the image segmentation technology has always attracted high attention and evolved over 1000 algorithms. However, there has been no fit-for-all algorithm of image segmentation so far, just proposing the method of segmentation relevant to specific issues [4].

With the development of computer image technology, segmentation method for insect images uses many new ways in image segmentation field for reference, such as smart segmentation using level sets, edge flow and in combination with shape, texture and color. However, the current segmentation technology remains the key to hindering wide application of insect images. The article [5] summarizes the existing segmentation algorithm and application for insect images. By overall analysis and evaluation as in [6, 7] articles, it is believed that the segmentation technology of insect images still possesses big room for research and improvement. Only by accurate segmentation of insect images, can it help improve accuracy in practical applications like calculation and species identification.

In practice, this paper mainly uses GrubCut algorithm [8], which is an important algorithm in OpenCV [9] used to delete image background. In use, users need to set ROI first and delete the

background within ROI then based on the areas outside ROI. However, in actual application, it is quite hard to set ROI in a unified manner due to variety of images to be processed and random distribution of ROI areas. It is also not realistic to set each image manually. Therefore, here we propose: set the outermost frame of image as background and those within the frame as ROI first so as to delete the background initially, screen the image after processing to keep the largest area in the image, calculate the position of selected area and reset the ROI to delete background, repeat the previous steps until the kept area in the image does not change any more, then we get the image finally without background. However, just as indicated by the above-mentioned part, because the image backgrounds are varied and complicated, there has not been a segmentation method yet that can completely and accurately delete the background area. Instead, efforts can be made to ensure that as much of background image as possible is deleted while ROI will not be deleted mistakenly.

Image Feature Extraction

The common image features include color, texture, shape and spatial relations, etc. In this system, color feature and textural feature are mainly adopted with no other features being used.

Color Feature

Color feature is a kind of full-scale feature description that is most commonly used in image analysis and processing technology. It includes many description methods like color histogram, color set and color matrix. The color histogram as a widely used feature in many image retrieval system, can briefly describe the overall distribution of colors in a picture, i.e. the share of different colors in the whole image, in particular fit to describe those images hard for automatic partitioning and with no need to consider the spatial position of objects. Therefore, color histogram cannot describe the local distribution of color in the image or the objects in the image either.

Currently the most common color histogram is grey level histogram, whose statistics come from H and S channels selected in the HSV color space. In this system, RGB three-channel color histogram is adopted, which combines three histograms on the R, G and B channels into a 3D histogram for comparison.

Textural Feature

Textural features describe repeated distribution of the same or similar areas in the image, which can reflect the image information like direction, variation range and intervals, and can be used effectively to calculate the ranking rule or local feature of images.

The common method to extract textural features of images is Grey Level Co-occurrence Matrix (GLCM), which calculates frequency of difference between two pixel grey levels on certain direction and on certain distance to get information of image on method, variation range and intervals so as to work out the textural features of images.

The GLCM of image on Line i and Row j with pixel distance at d and direction at θ is:

$$P = p(i, j, d, \theta) \tag{1}$$

Among which, $\theta = 0°, 45°, 90°, 135°$

Many textural feature values can be calculated through co-occurrence matrix, among which the following are widely used (in this system, the pixel distance d chooses the value of 1 and then textural feature values on four directions will be obtained) [10, 11]:

i. Entropy

$$\sum_i \sum_j P(i,j) * log\, P(i,j) \tag{2}$$

ii. Energy

$$\sum_i \sum_j P(i,j)^2 \tag{3}$$

iii. Inverse Difference Moment

$$\sum_i \sum_j \frac{P(i,j)}{1+(i-j)^2} \tag{4}$$

iv. Contrast

$$\sum_i \sum_j P(i,j) * (i-j)^2 \tag{5}$$

v. Correlation

$$\sum_i \sum_j P(i,j) \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x * \sigma_y} \quad (6)$$

Among which, $\mu_x = \sum_i \sum_j i * P(i,j)$, $\mu_y = \sum_i \sum_j j * P(i,j)$

$$\sigma_x = \sqrt{\sum_i \sum_j P(i,j) * (i-\mu_x)^2}, \sigma_y = \sqrt{\sum_i \sum_j P(i,j) * \left(j-\mu_y\right)^2}$$

vi. Maximum probability

$$\max\left(\sum_i \sum_j P(i,j)\right) \quad (7)$$

Image Feature Comparison

Different image features have different ways to compare, so we have the following specific methods for color histogram and textural features:

Color Histogram Comparison Algorithm

There are many methods for histogram comparison. Different methods will get different results based on different histogram data. In order to reduce error, here four comparison methods are used, i.e. correlation, chi-square, intersection and Bhattacharyya [12, 13]. The final image similarity will calculated based on the results of these four comparison methods at the same weight.

Correlation:

$$d_{correl}(H_1, H_2) = \frac{\sum_i H_1'(i) * H_2'(i)}{\sqrt{\sum_i H_1'^2(i) * \sum_i H_2'^2(i)}} \quad (8)$$

$H_k'(i) = H_k(i) - (1/N) \sum_j H_k(j)$ and N equals the number of bins in the histogram.

Chi-square:

$$d_{chi-square}(H_1, H_2) = \sum_i \frac{\left(H_1(i) - H_2(i)\right)^2}{H_1(i)} \quad (9)$$

Intersection:

$$d_{interection}(H_1, H_2) = \sum_i \min\left(H_1(i), H_2(i)\right) \quad (10)$$

Bhattacharyya distance:

$$d_{Bhattacharyya}(H_1, H_2) = \sqrt{1 - \sum_i \frac{\sqrt{H_1(i) * H_2(i)}}{\sum_i H_1(i) * \sum_i H_2(i)}} \quad (11)$$

Comparison Algorithm of Image Textural Features

As for textural feature value calculated by GLCM, mean value and variance of each feature on four directions is calculated to get feature vector of image. Then by working out difference in feature vector between two images, the comparison result of two images will be obtained.

The calculation process is Eq. 12.

$$d(V_2, V_2) = 1 - \frac{\sqrt{\sum_N \left(\frac{V_1(n) - V_2(n)}{\max(V_1(n), V_2(n))}\right)^2}}{N} \quad (12)$$

Among which, N is the length of feature vector, which is double of the feature value.

Image Clustering Algorithm

The clustering algorithm as a component of data mining, is a common tool of data analysis, whose aim is to classify sets of massive data into several types so that there is a maximum degree of similarity between data of each type and a largest degree of difference between data of different types. The clustering has arisen many classic algorithm as it has evolved to the current state, such as partitioning method, hierarchical method, density-based method, grid-based method and model-based method and so on [14]. It can be known through the above-mentioned introduction, the clustering algorithm and the classification algorithm differ significantly. The clustering is not clear about the number of types, which means 10 or 100 types are both possible, only depending on some similar conditions to cluster. Of course, there is also the clustering algorithm with user-defined types, but the numbers of types are not easy to define. Classification has the number of types already known and is also clear about the features of types. Then those unknown parts will be categorized into a

certain type based on some rules. Obviously, it is impossible to identify the number of species in the insect image in advance, and neither is possible to set the species classification rules in advance, so the clustering algorithm is needed for species classification.

The insect image features analyzed in this paper needs to set how to distinguish similarity of two images. We use the following formula: color histogram and image texture are given the same weight to calculate distance between two images by using the Euclid's distance. Then DBSCAN [15] algorithm will conduct clustering analysis on all the images, which involves two parameters. If the parameters are set improperly, then the clustering result will be affected greatly, leading to over-large (many different insect species are classified into one type) or over-small cluster (insect images of the same species are classified into different types). In practice, we can have numbers of tests to select proper parameters to get a better result.

Analysis and Summary

As far as the testing result is concerned, DBSCAN algorithm analysis basically matches expectation for species classification, which cut the manual workload evidently. Of course, there are still some errors, which may involve the following reasons:

1. The image background cannot be deleted accurately in an automatic way. In some cases, certain backgrounds will remain, such as the bees collecting honey;
2. We only use color histogram and textural features, so it is not sufficient to distinguish those species with subtle difference. In fact, something like shape feature can also be adopted;
3. Impact of parameter setting in DBSCAN.

In the follow-up work, we will continue to improve the method by centering on the above factors so as to further improve accuracy. The automatic classification algorithm, in essence, aims to identify insects in a shortest time and an accurate way. During this process, we will gradually set up the insect image feature database as the basis for insect identification, which will not only classify the insect images of the same species into one category, but also can judge the name of insect species and related information.

References

[1] Karunakaran C, Jayas D S, White N D G. Identification of Wheat Kernels damaged by the Red Flour Beetle using X-ray Images [J]. Biosystems Engineering, 2004, 87(3):1-8.

[2] Qiu D, Zhang H, Chen T, et al. Application of fuzzy recognition technique in stored-grain pests detection [J]. System Sciemces & Comprehensive Studies in Agriculture, 2002, 18(2):122-125.

[3] Yu XinWen, Shen Zuorui, Gao Lingwang, Li Zhihong. Feature measuring and extraction for digital image of insects. Journal of China Agricultural University 2003, 8(3):47-50

[4] Ouyang X Y, Zhao N N, Song L, et al. Survey on Image Segmentation [J]. Journal of Anshan Institute of Iron & Steel Technology, 2002.

[5] Wang J N, Li-Qiang J I. Methods of insect image segmentation and their application [J]. Acta Entomologica Sinica, 2011, 54(2):211-217.

[6] Yu, Xinwen, and Z. Shen. Segmentation Technology for Digital Image of Insects [J]. Transactions of the Chinese Society of Agricultural Engineering 2001.

[7] Qi-Xiang Y E, Gao W, Wang W Q, et al. A Color Image Segmentation Algorithm by Using Color and Spatial Information [J]. Journal of Software, 2004, 15(4):522-530.

[8] Rother C, Kolmogorov V, Blake A. Grabcut: Interactive foreground extraction using iterated graph cuts [C]//ACM Transactions on Graphics (TOG). ACM, 2004 23 (3) 309-314.

[9] OpenCV, open source computer vision, http://opencv.org/.

[10]HARALICK R M, SHANMUGAN K, DINSTEIN I. Texture features for image classification [J]. IEEE Trans on Systems, Man and Cybernetics, 1973, 3(6):610-621.

[11] CLAUSI D A. Texture segmentation of SAR sea ice imagery [D]. Waterloo: University of Wateloo, 1996.

[12] Gary Bradski, Adrian Kaebler. Learning OpenCV [M]. United States of America: O'Reilly Media, Inc., 2008. 201-203.

[13] OpenCV API Reference, http://docs.opencv.org/2.4.11/modules/imgproc/doc/histograms.html

[14] J. Han, M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, CA 2000.

[15] M. Ester, H.-P. Kriegel, J. Sander, X. Xu. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noises. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, 1996, pp. 226-231.